North American Climate in CMIP5 Experiments.

Part I: Evaluation of Historical Simulations of Continental and Regional Climatology

Justin Sheffield, Andrew Barrett, Brian Colle, D. Nelun Fernando, Rong Fu, Kerrie L. Geil, Qi Hu, Jim Kinter, Sanjiv Kumar, Baird Langenbrunner, Kelly Lombardo, Lindsey N. Long, Eric Maloney, Annarita Mariotti, Joyce E. Meyerson, Kingtse C. Mo, J. David Neelin, Zaitao Pan, Tong Ren, Alfredo Ruiz-Barradas, Yolande L. Serra, Anji Seth, Jeanne M. Thibeault, Julienne C. Stroeve, Ze Yang, Lei Yin

Justin Sheffield, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ

Brian Colle, Kelly Lombardo, School of Marine and Atmospheric Sciences, Stony Brook University - SUNY

Rong Fu, Lei Yin, Ze Yang, Tong Ren, Jackson School of Geosciences, University of Texas at Austin, TX

D. Nelun Fernando, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX and Visiting Scientist Programs, University Corporation for Atmospheric Research, Boulder, Colorado, and Surface Water Resource Division, Texas Water Development Board, Austin, TX

Kerrie L. Geil, Department of Atmospheric Sciences, University of Arizona, Tucson, AZ

Qu Hu, School of Natural Resources and Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, Lincoln, NE

Sanjiv Kumar, Jim Kinter, Center for Ocean-Land-Atmosphere Studies, Calverton, MD

Baird Langenbrunner, Joyce E. Meyerson, J. David Neelin, Department of Atmospheric and Oceanic Sciences, University of California Los Angeles

Lindsey N. Long, Wyle Science, Technology and Engineering, College Park, MD, and Climate Prediction Center/NCEP/NWS/NOAA, College Park, MD

Eric D. Maloney, Department of Atmospheric Science, Colorado State University, Fort Collins, CO

Annarita Mariotti, National Oceanic and Atmospheric Administration, Office of Oceanic and Atmospheric Research (NOAA/OAR), Silver Spring, MD

Kingtse C. Mo, Climate Prediction Center/NCEP/NWS/NOAA, College Park, MD

Zaitao Pan, Saint Louis University, St. Louis, MO

Alfredo Ruiz-Barradas, Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD

Yolande L. Serra, Department of Atmospheric Sciences, University of Arizona, Tucson, AZ

Anji Seth and Jeanne M. Thibeault, Department of Geography, University of Connecticut, Storrs, CT

Andrew Barrett, Julienne C Stroeve, National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO

Journal of Climate

Submitted on July 30, 2012

Revised February 3, 2013

*Corresponding author address: Justin Sheffield, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, 08540. Email: justin@princeton.edu

Abstract

This is the first part of a three-part paper on North American climate in CMIP5 that evaluates the historical simulations of continental and regional climatology with a focus on a core set of seventeen models. We evaluate the models for a set of basic surface climate and hydrological variables and their extremes for the continent. This is supplemented by evaluations for selected regional climate processes relevant to North American climate, including cool season western Atlantic cyclones, the north American monsoon, the US Great Plains low level jet, and Arctic sea ice. In general, the multi-model ensemble mean represents the observed spatial patterns of basic climate and hydrological variables but with large variability across models and regions in the magnitude and sign of errors. No single model stands out as being particularly better or worse across all analyses, although some models consistently outperform the others for certain variables across most regions and seasons, and higher resolution models tend to perform better for regional processes. The CMIP5 multi-model ensemble show a slight improvement relative to CMIP3 models in representing basic climate variables, in terms of the mean and spread, although performance has decreased for some models. Improvements in CMIP5 model performance are noticeable for some regional climate processes analyzed, such as the timing of the North American monsoon. The results of this paper have implications for the robustness of future projections of climate and its associated impacts, which are examined in the third part of the paper.

1. Introduction

This is the first part of a three-part paper on the Climate Model Intercomparison Project phase 5 (CMIP5; Taylor et al., 2012) model simulations for North America. The first two papers evaluate the CMIP5 models in their ability to replicate the observed features of North American continental and regional climate, and related climate processes for the recent past. This first part evaluates the models in terms of continental and regional climatology and the second part (Sheffield et al. 2013) evaluates intra-seasonal to decadal variability. The third part (Maloney et al., 2013) describes the projected changes for the 21st century.

The CMIP5 provides an unprecedented collection of climate model output data for the assessment of future climate projections as well as evaluations of climate models for contemporary climate, the attribution of observed climate change and improved understanding of climate processes and feedbacks. As such, these data feed into the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5), and other global, regional and national assessments. The goal of this study is to provide a broad evaluation of CMIP5 models in their depiction of North American climate and associated processes. The set of climate features and processes examined in this first part were chosen to cover the climatology of basic surface climate and hydrological variables and their extremes at daily to seasonal time scales, as well as selected climate features that have regional importance. The second part of this study (Sheffield et al. 2013) covers aspects of climate variability, such as intra-seasonal variability in the tropical Pacific, the El Niño Southern Oscillation (ENSO) and the Atlantic Multidecadal Oscillation that play major roles in driving North American climate variability. This study draws from individual work by investigators within the CMIP5 Task Force of the US National Oceanic and Atmospheric Administration (NOAA) Modeling Analysis and Prediction Program (MAPP). This is part of a Journal of Climate special collection on North America in CMIP5 models and we draw from individual papers within the special collection, which provide detailed analysis of some of the climate features examined here.

We begin in Section 2 by describing the CMIP5, providing an overview of the models analyzed, the historical simulations and the general methodology for evaluating the models. We focus on a core set of 17 CMIP5 models that represent a large set of climate centers and model types, and synthesize model performance across all analyses for this core set. Details of the observational datasets to which the climate models are compared are also given in this section. The next two sections focus on different aspects of North American climate and surface processes. Section 3 begins with an overview of climate model depictions of continental climate, including seasonal precipitation, air temperature, sea surface temperatures, and atmospheric and surface water budgets. Section 4 evaluates the model simulations of extremes of temperature and surface hydrology, and temperature-based biophysical indicators such as growing season length. The next section 5 focuses on regional climate features such as north Atlantic winter storms, the Great Plains low level jet, and Arctic sea ice. The results are synthesized in Section 6 and compared to results from CMIP3 models for selected variables.

2. CMIP5 Models and Simulations

2.1. CMIP5 Models

We use data from multiple model simulations of the "historical" scenario from the CMIP5 database. The scenarios are described in more detail below. The CMIP5 experiments were carried out by 20 modeling groups representing more than 50 climate models with the aim of further understanding past and future climate change in key areas of uncertainty (Taylor et al., 2012). In particular, experiments focus on understanding model differences in clouds and carbon feedbacks, quantifying decadal climate predictability and why models give different answers when driven by the same forcings. The CMIP5 builds on the previous phase (CMIP3) experiments in several ways. Firstly a greater number of modeling centers and models have participated. Secondly, the models generally run at higher spatial resolution with some models being more comprehensive in terms of the processes that they represent, therefore hopefully resulting in better skill in representing current climate conditions and reducing uncertainty in future projections. Table 1 provides an overview of the models used.

To provide a consistent evaluation across the various analyses, we focus on a core set of 17 models, which are highlighted in the table by asterisks. The core set was chosen to span a diverse set of modeling centers and model types (coupled atmospheric-ocean models (AOGCM) and Earth system models (ESM)), and includes an AOGCM and ESM from the same modeling center for three centers (GFDL, Hadley Center, AORI/NIES/JAMSTEC). The set was restricted by data availability and processing constraints, and so for some analyses (in particular those requiring high temporal resolution data) a smaller subset of the core models was analyzed. When data for non-core models were available, these were also evaluated for some analyses and the results are highlighted if they showed better (or particularly poor) performance. The specific models used for each individual analysis are provided within the results section where appropriate.

2.2. Overview of Methods

Data from the "historical" CMIP5 scenarios are evaluated in this study. The "historical" simulations are run in coupled atmosphere-ocean mode forced by historical estimates of changes in atmospheric composition from natural and anthropogenic sources, volcanoes, greenhouse gases and aerosols, as well as changes in solar output and land cover. For certain basic climate variables we also analyze model simulations from the CMIP3 that provided the underlying climate model data to the fourth assessment report (AR4) of the IPCC. Several models have contributed to both the CMIP3 and CMIP5 experiments, either for the same version of the model, or for a newer version, and this allows a direct evaluation of changes in skill in individual models as well as the model ensemble.

Historical scenario simulations were carried out for the period from the start of the industrial revolution to near present: 1850-2005. Our evaluations are generally carried out for the most recent 30 years, depending on the type of analysis and the availability of observations. For some analyses the only, or best available, data are from satellite remote sensing which restricts the analysis to the satellite period, which is generally from 1979 onwards. For other analyses, multiple observational datasets are used to represent the uncertainty in the observations. An overview of the observational datasets used in the evaluations is given in Table 2, categorized by variable. Further details of these datasets and any data processing are given in the relevant sub-sections and figure captions. Where the comparisons go beyond 2005 (e.g. 1979-2008), model data from the RCP8.5 future projection scenario simulation are appended to the model historical time series. Most of the models have multiple ensemble members and in general we use the first ensemble member. In some cases, the results for multiple ensembles are averaged where appropriate or used to assess the variability across ensemble members. Results are generally shown for the multi-model ensemble (MME) mean and for the individual models using performance metrics that quantify the errors relative to the observations.

3. Continental Seasonal Climate

We begin by evaluating the seasonal climatologies of basic climate variables: precipitation, near surface air temperature, sea surface temperature (SST), and atmosphere-land water budgets.

3.1. Seasonal Precipitation Climatology

Figure 1 shows the model precipitation climatology and GPCP (Adler et al., 2003) observations for December-February (DJF) and June-August (JJA) for 1979-2005. Table 3 shows the seasonal biases in precipitation for North America, the US and six regions. Most of the models do reasonably well in producing essential large-scale precipitation features and the bias in the MME mean seasonal precipitation over North America is about 12% and -1% for DJF and JJA, respectively. However, there are substantial differences among the models, and with observations at the regional scale (Table 3) and generally an overestimation of precipitation in more humid and cooler regions, and underestimation in drier regions. For the winter season (Fig. 1, left), the Pacific storm track is very reasonably placed in latitude as it approaches the coast. One important aspect of this, the angle of the storm track as it bends northward approaching the coast from roughly Hawaii to Central California, is well reproduced in the models. The intensity of the storm tracks off the West Coast compares reasonably well to the GPCP product shown here. The model coastal rainfall is not quite intense enough at the coast and spreads slightly too far inland, as might be expected for the typical model resolution which does not fully resolve mountain ranges and may help explain the overestimation by all models for western North America (WNA, see Table 3 for region definitions). The east coast storm tracks are well placed in DJF (see section 5.1 on winter-time extra-tropical cyclones) and the multi-model ensemble mean does a good job in replicating the Eastern Pacific Inter-Tropical Convergence Zone (ITCZ), although northern Mexico receives too much rainfall. Figures 1c provides a model by model view of these features using the 3 mm/day contour for each model to provide an outline of the major precipitation features. If the models were perfect, all contours would lie exactly along the boundary of the shaded observations. Taking into account the high latitude precipitation excess in the Pacific storm track, individual models do quite well at reproducing each of the main features of the DJF climatology, including the arrival point at the North American West Coast of the southern edge of the Pacific storm track. Only a few models exhibit the ITCZ extension feature that accounts for the northern Mexico precipitation excess.

For the summer season (JJA; Fig. 1, right), the ITCZ and the Mexican monsoon are reasonably well simulated in terms of position (see section 5.4 on the North American monsoon), although the precipitation magnitude in parts of the Caribbean is underestimated relative to GPCP. The East Coast storm track in the multi-model ensemble mean is too spread out and less coherent than observed. This is due to substantial differences in the placement of these storm tracks in the individual models (Fig. 1d). The majority of the models exhibit excessive precipitation in at least some part of the continental interior. While the bulk of the models do reasonably well at the poleward extension of the monsoon over Central America, Mexico and the Inter-Americas Seas region, a few models underestimate this extent, putting a split between the poleward extension of the monsoon feature and the start of the East Coast storm track. Overall, the models underestimate JJA precipitation over the Central America (including Mexico), and Central North America regions (Table 3).

3.2 Seasonal Surface Air Temperature Climatology

Figure 2 compares the model simulated surface air temperature climatology to the observation estimates from NCEP-DOE Reanalysis 2 and the CRU TS3.0 station-based analysis, here both shown interpolated to the same 2.5° grid as the models. The MME mean compares well to the observations in most respects. Differences from both observational estimates are less than or on the order of 1°C over most of the continent except for certain regions (see Table 4). The multi-model ensemble mean is cooler than both data sets over northern Mexico in DJF. In high latitudes, differences between the observational estimates are large enough that the error patterns in Fig. 2f and Fig. 2g differ substantially, especially in DJF (NCEP-DOE is also slightly warmer than the North American Regional Reanalysis, not shown, in this region and season). Beyond the overall simulation of the north-south temperature gradient and seasonal evolution, certain regional features are well represented. In JJA, this includes the regions of temperatures exceeding 30°C over Texas and near the Gulf of California, and the extent of temperatures above 10°C, including the northward extension of this region into the Canadian prairies. Individual model surface air temperature climatologies, shown in the supplementary material (Figs. S1 and S2) and in terms of biases in Table 4, exhibit substantial regional scatter, including excessive northward extent of the region above 30°C through the Great Plains in three of the models (CanESM2, CSIRO-Mk3-6-0 and FGOALS-s2). In DJF, the multi-model ensemble mean does a good job of representing the 0°C contour, while the 10°C contour extends slightly too far south, yielding slightly cool temperatures over Mexico, with 15 out of 18 models with cold biases over the broader CAM region (Table 4). The wintertime cold bias relative to both observational estimates in very high latitudes is more pronounced in certain models such as HadGEM2-ES, which is biased low by -7.0 and -5.1 ^{o}C over the ALA and NEC regions, respectively (Table 4). The inter-model scatter in surface temperature simulations is summarized in Fig. 2d for DJF and Fig. 2j for JJA using the inter-model standard deviation of the ensemble (i.e., the standard deviation at each gridpoint among the 18 model climatologies seen in Figs. S1 and S2). For DJF, the inter-model standard deviation is less than 2.5°C through most of the contiguous US but increases toward high latitudes, exceeding 3.5°C over much of the area north of 60°N. In JJA, there is a region of high inter-model standard deviation, exceeding 3.5°C, roughly in the Great Plains region in the northern US and southern Canada. This is a region with fairly high precipitation uncertainty in JJA (Fig. 1f), and changes in surface temperature in this region have been linked to factors affecting soil moisture, including pre-season snowmelt (e.g., Hall et al. 2008), so this may be a suitable target for further study to reduce model uncertainty.

3.3. Seasonal Sea Surface Temperature

The annual cycle of sea surface temperature (SST) is shown in Figure 3 as winter-to-spring (December-May) and summer-to-fall (June-November) means. We also show precipitation over land, which is generally associated with SST variations in adjoining ocean regions. Maps for individual models are shown in supplementary Figures S4 and S5. The Western Hemisphere Warm Pool (WHWP), where temperatures are equal or larger than 28.5°C, usually is absent from December to February, and appears in the Pacific from March to May, while it is present in the Caribbean and Gulf of Mexico from June to November (Wang and Enfield, 2001). The cooler part of the year is characterized by the small extension of SST in excess of 27°C and a suggestion of a cold tongue in the eastern equatorial Pacific, while during the warmer part of the year the extension of SSTs in excess of 27°C is maximum and the cold tongue is well defined over the eastern Pacific. High precipitation along the Mexican coasts, Central America, the Caribbean Islands and the central-eastern US are associated with the warm tropical SSTs during the warm half of the year. A decrease in the regional precipitation south of the equator is also evident in this warm half of the year.

The MME mean shows the observed change in SST from cold to warm around the WHWP region, however the warm pool is absent over the Caribbean and Gulf of Mexico region. The change in precipitation from the cold to the warm parts of the year is represented by the MME mean including the increase in precipitation over central US and

Mexico as well as the decrease south of the equator. The eastern Pacific in the models is slightly cooler than observations in the cold part of the year but not in the form of weak cold tongue from the Peruvian coast but rather as a confined equatorial cooling away from the coast. The cold tongue along the eastern equatorial Pacific and along the coast of Peru during the warmer part of the year is reasonably represented by the MME mean although its extension is farther to the west. Differences with observations of the multi-model mean indicate cool SST biases over the Pacific and Intra-American seas parts of the WHWP in all models in both the cold and warm parts of the year. Warm biases are evident close to the coasts of northeastern US, western Mexico and Peru. Precipitation biases indicate a wet/dry bias to the west/east of ~97°W northward of 15°N over Mexico and the US during both parts of the year (as well as the intense and extensive dry bias over South America to the east of the Andes); the cold bias over the Intra-Americas sea and the dry bias over the Great Plains in the US suggests a link between the two, considering the former is a great source of moisture for the latter.

Spatial statistics for the mean annual SSTs are summarized in Table 5 for the individual CMIP5 models, and the MME mean. The spatial correlations are > 0.9 for all models, and are not able to quantitatively distinguish the performance of the models. The MME mean maximizes the spatial correlation (0.97) and minimizes the RMSE (0.77°C), but not the bias (-0.54°C). Eight of the models have RMSE values less than 1°C, and the largest biases (> 1.3°C) are for CSIRO-MK3.6, HadCM3, INMCM4, IPSL-CM5A-LR, and MIROC-ESM. The biases, except for INMCM4 and CCSM4, are negative, with the smallest bias for INMCM4, and the largest for CSIRO-MK3.6.

3.4. Seasonal Atmospheric and Land Water Budgets

We next evaluate the climatologies of the atmospheric and land water budgets. Seasonal changes in atmospheric water content are relatively small compared to the moisture fluxes and so we focus on the latter. Variations in moisture divergence are generally correlated with seasonal precipitation and so may help explain biases in model precipitation. The vertically integrated moisture transport (vectors) and its divergence (contours) are shown in Fig. 4 for five CMIP5 models (the number of models was limited by the availability of high temporal resolution model data required to calculated the moisture fluxes) and observational estimates from the 20CR for mean JJA and DJF for 1981-2000. In summer, the 20CR shows southerly transport from the North Atlantic anticyclone that splits into two distinct branches: one flanking the Atlantic seaboard with large scale convergence off the east coast and a second branch of moisture flows into the interior central plains which is associated with convergence over the Rocky Mountains. The western U.S. is dominated by divergence associated with the northerly component of the North Pacific anticyclone. The five models show the two branches of moisture transport, with associated convergence off the east coast and divergence in the plains, albeit weaker. They also simulate the divergence in much of the west, but they do not simulate the strong convergence over the Rockies and Mexican Plateau as seen in 20CR, which is associated with the low bias in precipitation over these regions (Table 3; mean biases for the five models shown here are -19.8% and -31.5% for CNA and CAM regions, respectively). Spatial correlations for divergence in the North American region range from 0.08 to 0.42, with MIROC5 and CNRM-CM5 performing the best out of the five models according to this measure (Table 6). In winter, the 20CR shows a more zonal transport than during summer, with weaker flow around the subtropical anticyclones and moisture convergence across much of the continent. The models represent both the moisture transport and divergence patterns well including the stronger convergence in the Pacific Northwest and northern California and divergence in southern California, although the magnitude of divergence is too strong along the coasts, most notably for the CCSM4 and CNRM-CM5 models, and precipitation over the western North America is overestimated by all five models examined here (WNA and ALA regions, Table 3), and especially for the CCSM4. The improvement in winter over summer for the whole domain is evident in the spatial correlations, which range between 0.60 and 0.76 for winter, with a different set of models performing better than in summer (CanESM2, CCSM4, and CNRM-CM5 – Table 6).

Evaluations of the model simulated terrestrial water budget are shown in Figures 5 and 6 against the off-line land surface model (LSM) simulations. Fig. 5 shows the regional mean seasonal cycles of the components of the land surface water budget (precipitation, evapotranspiration, runoff, change in water storage). In reality, water storage includes soil moisture, surface water such as lakes, reservoirs, and wetlands, groundwater and snowpack, but, in general, the climate models only simulate the soil moisture and snowpack components. Figure 5 also separates out the snow component of the water budget in terms of the snow water equivalent (SWE). Most models have a reasonable seasonal cycle of precipitation and evapotranspiration but tend to overestimate precipitation in the more humid and cooler regions (WNA, ENA, ALA, NEC) as noted previously and overestimate evapotranspiration throughout the year and especially in the cooler months. Runoff is generally underestimated, particularly in the central and eastern

North American regions (CNA and ENA) and in northeast Canada (NEC) and central America (CAM). It also peaks earlier in the spring in some models (that can be linked to a shortened snow season; see below), although the models generally replicate the spatial variability in annual total runoff (Figure 6 and Figure S6 in the supplementary material). The majority of models overestimate total runoff over dry regions and high latitudes, particularly for the Pacific Northwest and Newfoundland. SWE is generally overestimated by the multi-model ensemble for western North America, underestimated in the east and overestimated in the Alaskan/Western Canada region, which are a reflection of the precipitation biases. These biases are also reflected in the change in storage, particularly for the Alaska region where many of the models show a large negative change during late spring melt due to overestimation of SWE.

Figure 6 (Figure S6 for individual models) also shows the runoff ratio (runoff divided by precipitation) over North America, which indicates the production of water at the land surface that is subsequently potentially available as water resources. The remaining precipitation is partitioned into evapotranspiration (assuming that storage does not change much over long time periods). Overall the MME mean replicates the spatial pattern from the observational estimate with higher ratios in humid and cooler regions, and lower ratios in dry regions. However, the MME mean overestimates the ratios in humid and cooler regions, especially in Alaska, western and northern Canada, and underestimates the ratios in dry regions (Table 7). For North America overall, the models overestimate the ratios. The biases in runoff are better explained by biases in runoff ratios rather than biases in precipitation (not shown), especially in higher latitudes, highlighting the importance of the land surface schemes in the climate models and whether they are able to realistically partition precipitation into runoff and evapotranspiration, and accumulate and melt snow.

4. Continental Extremes and Biophysical Indicators

This next section examines the performance of the models in representing observed temperature and hydrological extremes. We first focus on temperature extremes and temperature dependent biophysical indicators, and then persistent seasonal hydrological extremes for precipitation and soil moisture. Regional extremes in temperature and precipitation are evaluated in Section 5.

4.1. Temperature Extremes and Biophysical Indicators

Temperature extremes have important consequences for many sectors including human health, ecosystem function, and agricultural production. We evaluate the models' ability to replicate the observed spatial distribution over North America of the frequency of extremes (Figure 7) for the number of summer days with maximum temperature (Tmax) > 25^{o}C and the number of frost days with minimum temperature (Tmin) < 0^{o}C (Frich et al., 2002) and a set of biophysical indicators related to temperature: spring and fall freeze dates and growing season length. We define the growing season length following Schwartz et al. (2006) which is the number of days between the last spring freeze of the year and the first hard freeze of the autumn in the same year. A hard freeze is defined as when the daily minimum temperature drops below -2^{o}C.

Overall, the models tend to underestimate the number of summer days by about 18 days over North America (Table 8), with regional underestimation of over 50 days in the western US and Mexico, and parts of the eastern US, but otherwise are within 20 days of the observations for most other regions. Several models (CanESM4, CCSM4, CNRM-CM5, MIROC, MIRCO-ESM) overestimate the number of summer days from the northeastern US up to the Canadian Northern Territories, but tend to have smaller underestimation in the western US and Mexico (see supplementary Figure S7). Nearly all other models have low biases of up to 50 days in these drier regions, which, at least for the western US, may be related to overestimation of precipitation and evapotranspiration (as shown in Section 3.4) and thus a reduction in sensible heating of the atmosphere. Several models have small biases for North America as a whole (Table 8), but often because large regional biases cancel out, and only the BCC-CSM1-1, CSIRO-Mk3-6-0 and HadGEM2-ES models have reasonably low biases (< 30 days) across all regions. The first two of these models also have relatively low runoff ratio biases for the western and central North American regions (WNA, CNA) (HadGEM2-ES was not evaluated for surface hydrology) suggesting that their simulation of warm summer days is not impeded by biases in the surface energy budget. The number of frost days are better simulated in terms of overall MME mean bias (-2.8 days) but there is a positive bias for most models across the Canadian Rockies and down into the US Rockies for most models (see supplementary Figure S8). Some of the models are biased low in the central US by over 50 days. Models with the least bias in frost days also tend to be the least biased models for summer days, but again many of the regional biases cancel out for the North America values.

The models do reasonably well at depicting the spatial distribution of growing season length (MME mean bias = -8.5 days over North America). The largest biases of between 30-50 days are in western Canada where the models underestimate and in the central US where they overestimate. The former is mainly because the last spring freeze is too late in western Canada and for the latter because of biases in both the last spring freeze (too early) and the first autumn freeze (too late). The INMCM4 model has the largest bias overall (-76 days), which is consistent over most of the continent (see Figure S9). The MIROC5 and MIROC-ESM models have the largest overestimations of 33 and 38 days, respectively, and these biases are also consistent over much of the continent.

4.2. Hydroclimate Extremes

We examine the ability of CMIP5 models to simulate persistent drought and wet spells in terms of precipitation and soil moisture (SM). We focus on the US because of the availability of long-term estimates of SM from the NLDAS-US dataset. Meteorological drought and wet spells are characterized by the 6-month Standardized Precipitation Index (SPI6; McKee et al., 1993). Agricultural drought and wet spells are evaluated in terms of soil moisture percentiles (Mo, 2008). The record length, N_{total}, is defined as the total months from all ensemble simulations of a model or the total months of the observed data set. At each grid point, an extreme negative (positive) event is selected when the SPI6 index is below (above) -0.8 (0.8) for a dry (wet) event [*Svoboda et al.* 2002]. For SM percentiles, the threshold is 20% (80%) for a dry (wet) event. At each grid cell, the number of months that extreme events occur (N) is 20% of the record length by construct (N/N_{total} = 20%). Because a persistent drought event (wet event) usually means persistent dryness (wetness), a drought (wet) episode is selected when the index is below/above this threshold for 3 consecutive seasons (9 months) or longer. The frequency of occurrence of persistent drought or wet spells (FOC) is defined as: FOC= N_p/N, where N_p is the number of months that an extreme event persists for 9 months.

Figures 8 and 9 show the FOC averaged for persistent wet and dry events for SPI6 and SM, respectively for 15 of the core models (the GFDL-ESM2M and INMCM4 model datasets only had a single ensemble member and the total record is therefore too short for the analysis). The most noticeable feature is the east-west contrast of the FOC for both SPI6 and SM as driven by the gradient in precipitation amount and variability (Mo and Schemm, 2008). Persistent drought and wet spells are more likely to occur over the western interior region, while extreme events are less likely to persist over the eastern US and the west coast. The maxima of the FOC are located in two bands, one located over the mountains and one extending from Oregon to Texas (Fig. 8-a). Persistent events are also found over the Great Plains. The CanESM2, CCSM4 and MIROC5 models show the east-west contrast, although the magnitudes of FOC are too weak for the CanESM2 model. The center of maximum FOC for MIROC5 is too far south.

Table 8 shows the performance of the models in representing the east-west contrast in terms of a FOC index defined as the difference in the fraction of grid cells with FOC greater than a given threshold between the western (32-48^{o}N, 92-112^{o}W) and eastern (32-48^{o}N, 70-92^{o}W) regions. The thresholds are 0.2 for SPI and 0.3 for SM. The FOC index values for the CCSM4 (0.35) and MIROC5 (0.34) models are closest to the observations (0.37) for SPI6. The MPI-ESM-LR model also shows the east-west contrast with one maximum located over Utah and another over the Great Plains, but the second maximum is too spatially extensive. The MIROC-ESM, MRI-CGCM3, and NorEMS1-M models all show a band of maxima over the Southwest, but the FOC north of 35^{o}N is too weak. Other models such as CSIRO-Mk3.6.0, IPSL-CM5A-LR, CNRM-CM5, GISS-E2-R and GFDL-CM3 have the maxima located over the Gulf region, which is too far south. Finally, the HadCM3 and HadGEM2-ES (not shown) models do not have enough persistent events.

For SM (Figure 9), the FOC from the NLDAS-UW shows that persistent anomalies are located west of 90^{o}N over the western interior region, with a FOC index of 0.68. Many of the models, such as BCC-CSM1.1, HadCM3, and IPSL-CM5A-LR, do not have enough persistent events, and the CanESM2, GISS-E2-R, and MRI-CGCM3 models shift the maxima to the central US. The CCSM4, GFDL-CM3, and NorESM1-M models fail to replicate the east-west contrast because of their high FOC values throughout most of the US. The best model for SM is the MPI-ESM-LR with a FOC index of 0.62, because it represents the east-west contrast and also has realistic magnitudes. The CSIRO-Mk3.6.0 model also simulates the east-west contrast, but the maximum is located south of the NLDAS-UW analysis maximum.

5. Regional Climate Features

We next evaluate the CMIP5 models for a set of regional climate features that have important regional consequences, either directly such as extreme temperature and precipitation in the southern US, and the North American Monsoon, or indirectly such as western Atlantic cool season cyclones and the US Great Plains low level jet. The last analysis examines the simulation of Arctic sea ice, which is important locally but also has implications for North American climate and elsewhere (Francis and Vavrus, 2012).

5.1. Cool Season Western Atlantic Extratropical Cyclones

Extratropical cyclones can have major impacts (heavy snow, storm surge, winds, flooding) along the east coast of North America given the proximity of the western Atlantic storm track. The Hodges (1994; 1995) cyclone tracking scheme was implemented to track cyclones in 15 models (of which 12 were in the core set) for the cool seasons (November to March) for 1979-2004. The CFSR reanalysis was used to estimate observed cyclone tracks. Six-hourly mean sea-level pressure (MSLP) data were used to track the cyclones, since it was found that including 850-hPa vorticity tracking yielded too many cyclones. Since MSLP is strongly influenced by large spatial scales and strong background flows, a spectral bandpass filter was used to preprocess the data. Those wavelengths between 600 and 10,000 km were kept, and the MSLP pressure anomaly had to persist for at least 24 hours and move at least 1000 km. Colle et al. (2013) describes the details of the tracking approach and validation of the tracking procedure.

Figure 10 shows the cyclone density during the cool season for the CFSR, mean and spread of the 15 models (see the legend of Fig. 11 for a complete listing), and select models for eastern North America and the western and central North Atlantic. There is a maximum in cyclone density in the CFSR over the Great Lakes, the western Atlantic from east of the Carolinas northeastward to east of Canada, and just east of southern Greenland (Figure 10a). The largest maximum over the western Atlantic (6-7 cyclones per cool season per 50,000 km^2) is located along the northern boundary of the Gulf Stream current. The MME mean is able to realistically simulate the three separate maxima locations (Figure 10b), but the amplitude is 10-20% underpredicted. The cyclone density maximum over the western Atlantic does not conform to the boundary of the Gulf

Stream as much as observed. There is a large inter-model spread near the Gulf Stream, since some models are able to better simulate western Atlantic density amplitude, such as the CCSM4 and HadGEM2-CC (Figures 10e,f). However, the CCSM4 maximum is shifted a few hundred kilometers to the north.

The distribution of cyclone central pressures at their maximum intensity were also compared (Figure 11) between the CFSR, MME mean, and individual models for the dashed box region in Figure 10b. There is a peak in cyclone intensity in both the CFSR and MME mean around 900-1000 hPa, and there is large spread in the model intensity distribution by almost a factor of two. The ensemble mean realistically predicts the number of average strength to relatively weak cyclones; however, the intensity distribution is too narrow compared to the CFSR, especially for the deeper cyclones < 980 hPa.

Colle et al. (2013) verified the 15 models by calculating the spatial correlation and mean absolute errors of the cyclone track densities and central pressures. They ranked the models and showed that 6 of the 7 best models were the higher resolution models (top three: EC-Earth, MRI-CGM3, and CNRM-CM5), since many lower resolution models, such as GFDL-ESM2M (Figures 10d), underpredict the cyclone density and intensity. The MME mean calculated using the 12 core models has verification scores within 5% of those from all 15 models (not shown), so it is likely that using all 17 core models in the cyclone analysis would not have much impact on the results.

5.2. Northeast Cool Season Precipitation

We next examine regional precipitation in the highly populated northeast US, which is expected to increase in the future (Maloney et al., 2013). The focus is on the cool season, since extratropical cyclones provide much of the heavy precipitation in the northeast. 14 of the core models (listed in Figure 11; daily precipitation data were not available for three models) were evaluated for the cool seasons (November to March) of 1979-2004. The model daily precipitation was compared with the CPC-Unified daily precipitation at 0.5 degree and CMAP monthly precipitation at 2.5 degree resolution.

Figures 12a-c shows the seasonal average precipitation for the two observational analyses and the MME mean and spread. The heaviest precipitation (700-1000mm) is over the Gulf Stream, which is associated with the western Atlantic storm track. This maximum is well depicted in the multi-model mean, although it is underestimated by 50-200mm, and there is a moderate spread between models (100-200mm). The precipitation over the northeast US ranges from 375mm in the northwestern part to around 500mm at the coast. The finer resolution CPC-Unified analysis has more variability downstream of the Great Lakes (lake effect snow) as well as some terrain enhancements. The models cannot resolve these smaller scale precipitation features, but the MME mean realistically represents the north to south variation. However, the MME mean overestimates precipitation by 25-75 mm (5-20%) over northern parts. Much of this overestimation is for thresholds greater than 5 mm day^{-1} over land. (Fig. 12d). The seasonal precipitation MME spread over the northeast is 100-150mm (25-40%), and much of this spread is reflected in the higher (> 10 mm day^{-1}) thresholds, with the BCC-CMS1-1 simulating less than the CPC Unified analysis, and a cluster of models, such as the INMCM4 and MIROC5, having many more heavy precipitation events than observed.

The model precipitation was verified against the CPC-Unified analysis for the black box region over the northeast US shown in Fig. 12b, and the models ranked in terms of their mean absolute errors (MAE) (Table 10). The MME mean has the lowest MAE. There is little relationship with resolution, since some relatively higher resolution models (e.g., MIROC5 and MRI-CGCM3) perform worse than many other lower resolution models. Most models have a 5-15% high bias in this region. There is little correlation (~0.22) between the high biases in precipitation in this region and the cyclone overestimation along the US East coast, thus suggesting the cyclone biases are coming from other processes than diabatic heating errors from precipitation.

5.3. Extreme Temperature and Rainfall over the Southern US

The southern regions of the US are historically prone to extreme climate events such as extreme summer temperatures, flood and dry spells. Previous CMIP and US climate impact assessments (Karl et al. 2009) have projected a large increase of these extreme events over regions of the south (southwest (SW), south central (SC), southeast (SE)), especially for the SW and SC US. However, to what extent climate models can adequately represent the statistical distributions of these extreme events over these regions is still unclear. Figure 13 compares the model simulated precipitation and temperature with observations as Taylor diagrams for 1) the annual number of heavy precipitation days (precipitation > 10mm day^{-1}) and 2) the number of hot days (Tmax > 32^{o}C (90^{o}F)). The observations are derived from the GHCN daily Tmax and Tmin gauge data and the CPC US-Mexico daily gridded precipitation dataset. Results are shown for 15 models, 11 of which are core models, in terms of the spatial correlation with the observations and standard deviation normalized by the observations. Table 11 also shows the regional biases.

Overall, the spatial distribution of the number of heavy precipitation days is better simulated in the SW and SC than the SE, for which the spatial correlations are below 0.5, with many models having negative correlations. The normalized standard deviations are less than observed indicating that the models cannot capture the high spatial variability in this region. Part of the reason for this may be the severe underestimation of number of tropical cyclones (Sheffield et al., 2013), although other factors are likely involved such as the biases in summertime convective precipitation. For the SW and SC regions the models do reasonably well at replicating the spatial variation, although with some spread across models (correlation values of 0.56-0.91, and 0.59-0.97, for the SW and SC, respectively). The MME mean simulated number of heavy precipitation is biased slightly high for the SW (but note that the observed number of days, 8.5, is small), and low for the SC and SE. For individual models, the GISS-E2-R model has a large high bias in the SW and the CanESM2, GFDL-CM3, HadCM3, IPSL-CM5A-LR and MIROC5 models have large low biases (> 10 days) in the SC and SE. Several models do reasonably well for all regions (GFDL-ESM2G, GFDL-ESM2M, HadGEM2-CC, HadGEM2-ES, MIROC4h, MPI-ESM-LR and MRI-CGCM3) in terms of their biases.

The number of hot days (Tmax > 32^{o}C) are underestimated by the MME mean for all regions by between about 12-19 days, which is consistent with the underestimation of summer days (Tmax > 25^{o}C) shown for N. America in Figure 7. Again the performance of the models in terms of spatial patterns and variability, and regional bias is generally worse for the SE. Interestingly, the three Hadley Center models considered here

(HadCM3, HadGEM2-CC and HadGEM2-ES) have the lowest biases for the SW and SC (except for the CCSM4), and in the SE (except for the MIROC models). The models tend to overestimate the spatial variability in the SW and underestimate it the SE, and the spatial correlations for the SW > SC > SE. The MIROC4h, which is a very high-resolution model (0.56 degree grid), stands out for all regions and both variables as having high spatial correlation, and low bias for heavy precipitation days, although it generally has too high spatial variability relative to the observations.

5.4. North American Monsoon

The North American Monsoon (NAM) brings rainfall to southern Mexico in May, expanding northward to the southwest US by late June or early July. Monsoon rainfall accounts for roughly 50-70% of the annual totals in these regions (Douglas et al., 1993; Adams and Comrie, 1997), with the annual percentages decreasing northward where winter rains become increasingly important. The annual cycle of precipitation from the ITCZ through the NAM region is examined in Figure 14. The MME mean from the 17 core models (averaged for longitudes 102.5°-115°W for 1979-2005) replicates the northward migration of precipitation in the NAM region during the warm season, but is biased low. However, the MME mean precipitation begins later, ends later, and is stronger than the observed estimate from CMAP within the core monsoon region north of 20°N. Within the latitudes of the ITCZ (up to 12°N), the models strongly underestimate the precipitation and fail to show the northward migration from stronger precipitation in May south of 8°N to a maximum in July near 10°N. Instead, the models tend to place the spring maximum at 10°N and have a late build up and late demise at all latitudes of the

ITCZ through boreal summer. Table 12 shows the RMSE for individual models over the domain shown in Figure 14 and indicates that the CanESM2, HadCM3 and HadGEM2-ES models have the lowest errors (< 0.75 mm day^{-1}) and the BCC-CSM1-1, NorESM1-M and MRI-CGCM3 the highest (> 1.9 mm day^{-1}).

The seasonal cycle of monthly precipitation in the core NAM region of northwest Mexico (23.875°-28.875°N, 108.875°-104.875°W) is also examined in Table 13 and Figures 15 for the core models plus four other models. Our core domain is similar to that used by the North American Monsoon Experiment (NAME; Higgins et al. 2006) and related studies (e.g., Higgins and Gochis, 2007; Gutzler et al., 2009), but has been reduced in size to ensure consistency of the monsoon precipitation signal at each grid point. Following the methodology of Liang et al. (2008) for analysis of CMIP3 data, we calculate a phase and RMS error of each model's seasonal cycle, where the phase error is defined as the lag in months with the best correlation to the observations (Table 13). The observations used are the P-NOAA, which is a recently developed gauge-based dataset that is likely more accurate than the CMAP for this region. We additionally calculate each model's annual bias as a percentage of the mean monthly climatological P-NOAA value (1.66 mm day^{-1}). The seasonal cycles for models with small (lag=0), moderate (lag=1) and large (lag=2-4) phase errors are shown in Figure 15a-c. Figure 15d shows the MME mean for all phase errors, their spread and the observations.

Overall the small phase error models tend to overestimate rainfall in the core NAM region compared to the two observational data sets throughout the year, with the largest errors seen in fall, consistent with Figure 14. The overestimation of rainfall by the models beyond the end of the monsoon season is also apparent in the small and large phase error CMIP3 models (Liang et al. 2008). The similarity between the range of RMSE values (0.46-2.23 mm day^{-1}) in their study of CMIP3 models and that of the CMIP5 models in this analysis indicates that there has been no improvement in the magnitude of the simulated annual cycle of monthly precipitation, with the lowest and highest RMSE values having increased slightly since the previous generation of models. On the other hand, there does seem to be improvement in the timing of seasonal precipitation shifts, with 13 out of 21 (62%) CMIP5 models having a phase lag of zero months as compared to 6 out of 17 (35%) CMIP3 models in Liang et al. (2008). The top ranking models for phase, RMSE and bias shown in Table 13 (HadCM3, HadGEM2-ES, CNRM-CM5, CanESM2, HadGEM2-CC) are also the models with the highest spatial correlations of May-October 850hPa geopotential heights and winds when compared with the ERA-Interim (Geil et al. 2013). The HadCM3, HadGEM2-ES and CanESM2 also perform the best over the larger monsoon region (Table 12). Geil et al. (2013) find that the models that best represent the seasonal shift of the monsoon ridge and subtropical highs over the North Pacific and Atlantic tend to have the least trouble ending the monsoon, suggesting there is room for improvement over the region through an improved representation of the seasonal cycle in these large-scale features.

5.5. Great Plains Low Level Jet

An outstanding feature of the warm season (May-September) circulation in North America is the strong and channeled southerly low-level flows, or the Great Plains low-level jet (LLJ), from the Gulf of Mexico to the central US and the Midwest (Bonner and Paegle 1970; Mitchell et al. 1995). The LLJ emerges in early May in the transition of the circulation from the cold to the warm season. It reaches its maximum strength in June and July. After August, the jet weakens and disappears in September when the cold season circulation starts to set in. While many studies have examined specific processes associated with the LLJ (Blackadar, 1957; Wexler, 1961; Holton, 1967), such as its nocturnal peak in diurnal wind speed oscillation, as well as precipitation, the jet is a part of the seasonal circulation shaped primarily by the orographic configuration in North America, particularly the Rocky Mountain Plateau (e.g., Wexler, 1961). An important climatic role of the LLJ is transporting moisture from the Gulf of Mexico to the central and eastern US (Benton and Estoque, 1954; Rasmusson, 1967; Helfand and Schubert, 1995; Byerle and Paegle, 2003). Because the moisture is essential for development of precipitation, even though additional dynamic processes are required for the latter to happen (Veres and Hu 2013), correctly describing the LLJ and its seasonal cycle is critical for simulating and predicting warm season precipitation and climate in central North America.

Outputs from eight of the core models (CanESM2, CCSM4, CNRM-CM5, GFDL-ESM2M, HadGEM2-ES, MIROC5, MPI-ESM-LR, and MRI-CGCM3) were analyzed for their simulation of the LLJ. Figure 16 compares the spatial profile and seasonal cycle between the MME mean and the NCEP/NCAR reanalysis in terms of the summer 925hPa winds, the vertical structure of the summer meridional wind, and the seasonal cycle of the LLJ. While the overall features of the simulated LLJ compare well with the reanalysis results, several details differ. First of all the models produce a peak meridional wind around 925hPa whereas the reanalysis result peaks around 850hPa. This difference has little impact from the vertical resolution of the models and the reanalysis because they share the same vertical resolution below 500hPa. For a few models that have more model levels below 500hPa their vertical profile of the meridional wind shows a similar peak at 925hPa. The vertical extent of the LLJ is shallower than that shown in the reanalysis, as suggested by the differences in Fig. 16f, which may be related to the peak wind being at a lower level in the troposphere. Secondly, the simulated LLJ extends much further northwards in the Great Plains than the reanalysis (Fig. 16g,h,i). For the seasonal cycle, the models show strong southerly winds that persist from mid-May to near the end of July whereas the reanalysis shows that the LLJ weakens substantially in early July (Fig. 16i). While these detailed differences exist, the error statistics in Table 14 indicate that these eight models simulate the LLJ satisfactorily.

5.6. Arctic/Alaska Sea Ice

Since routine monitoring by satellites began in late October 1978, Arctic sea ice has declined in all calendar months (e.g. Serreze et al., 2007). Trends are largest at the end of the summer melt season in September with a current rate of decline through 2012 of -14.3% per decade. Regionally, summer ice losses have been pronounced in the Beaufort, Chukchi and East Siberian seas since 2002 causing a lengthening of the ice-free season. The presence of sea ice helps to protect Alaskan coastal regions from wind-driven waves and warm ocean water that can weaken frozen ground. As the sea ice has retreated further from coastal regions, and ice-free summer conditions are lasting for longer periods of time (in some regions by more than 2 months during the satellite data record), wind-driven waves, combined with permafrost thaw and warmer ocean temperatures, have led to rapid coastal erosion (Mars and Houseknecht, 2007; Jones, et al., 2009).

While the winter ice cover is not projected to disappear in the near future, all models that contributed to the IPCC 2007 report showed that as temperatures rise, the Arctic Ocean would eventually become ice-free in the summer (e.g. Stroeve et al., 2007). However, estimates differed widely, with some models suggesting a transition towards a seasonally ice-free Arctic may happen before 2050, and others, sometime after 2100. To reduce the spread some studies suggest using only models that are able to reproduce the historical sea ice extent (e.g. Overland et al., 2011; Wang and Overland, 2009).

Historical sea ice extent (1953-2005) from 26 models during September and March is presented as box and whisker plots (Figure 17), constructed from all ensemble members of all models, with the width of the box representing the number of ensemble members. Table 15 shows the biases for the individual models. Five climate models (CanESM2, EC-EARTH, GISS-E2-R, HadGEM2-AO and MIROC4h) have mean September extents that fall below the minimum observed value, with EC-EARTH, GISS-E2-R and CanESM2 having more than 75% of their extents below the minimum observed value. Three models (CSIRO-Mk3-1-0, FGOALS-s2, NorESM1-M) have more than 75% of their extents above the maximum observed value. Overall, 14 models have mean extents below the observed 1979-2005 mean September extent. During March, several models fall outside the observed range of extents, with 16 models having more than 75% of their extents outside the observed maximum and minimum values (8 above, 8 below). Six models essentially straddle the mean observed March sea ice extent.

Spatial maps of March and September CMIP5 sea ice thickness averaged from 1993 to 2005 are shown in Figure 18 together with thickness estimates from ERS1/2 (1993-2001; Laxon et al., 2003), ICESat (2003-2009; Kwok and Cunnigham, 2008) and

IceBridge (2009-2012; Kurtz et al., 2012). Table 15 shows the biases for the individual models relative to the ICESat data. While we do not expect the models to be in phase with the observed natural climate variability and therefore accurately represent the magnitude of the ICESat thickness fields, it is important to assess whether or not the models are able to reproduce the observed spatial distribution of ice thickness. Data from ICESat and IceBridge, as well as earlier radar altimetry missions (ERS1/2) and submarine tracks indicate that the thickest ice is located north of Greenland and the Canadian Archipelago (> 5m thick) where there is an onshore component of ice motion resulting in strong ridging. Thicknesses are smaller on the Eurasian side of the Arctic Ocean where there is persistent offshore motion of ice and divergence, leading to new ice growth in open water areas. Most models fail to show thin ice close to the Eurasian coast and thicker ice along the Canadian Arctic Archipelago and north coast of Greenland. Instead, many models show a ridge of thick ice that spans north of Greenland across the Lomonosov Ridge towards the East Siberian Shelf, with thinner ice in the Beaufort/Chukchi and the Kara/Barents seas. In large part this is explained in terms of biases in the distribution of surface winds; for example, if a model fails to produce a well-structured Beaufort Sea High, this will adversely affect the ice drift pattern and hence the thickness pattern. Nevertheless, when we compare mean thickness fields from IceBridge, ICESat and ERS1/2 with thickness fields from the CMIP5 models for the period 2000 to 2010, we find that for the Arctic Ocean as a whole, the thickness distributions from the models overlap with those from the satellite and airborne derived products. However, for the North American side of the Arctic Ocean model thicknesses tend be smaller than thicknesses estimated from derived products. This in part explains the low bias in September ice extent for some of the models, as thinner ice is more prone to melting out in summer. Models with extensively thick winter ice (e.g. NorESM1-M and MIROC5) on the other hand tend to overestimate the observed September ice extent.

6. Discussion and Conclusions

6.1. Synthesis of Model Performance

This study evaluates the CMIP5 models for a set of basic climate and surface hydrological variables for annual and seasonal means and extremes, and selected regional climate features. Evaluations of model performance are not straightforward because of the broad range of uses of climate model data (Glecker et al., 2008) and therefore there is not an accepted universal set of performance metrics. Issues relevant to performance are dependent on several elements including decadal variability, observational uncertainties, and that some models are tuned to certain processes, often at the expense of other aspects of climate. The performance metrics evaluated here are generally focused on basic climate variables and standard statistical measures such as bias, RMSE and spatial correlation. One of the strengths this study is the broad range of evaluations that test multiple aspects of the model simulations at various time and space scales, and for specific important regional features that we do not necessarily expect coarse resolution models to simulate well. Independently these metrics indicate much better performance by certain models relative to the ensemble, whilst some models have poor performance in that a feature is not simulated at all, such as lack of persistence in extreme hydrological events, or the errors are unacceptably large. However, it is not clear whether one model or set of models performs better than others for the full set of climate variables.

Figure 19 shows a summary ranking of model performance across all continental and US domain analyses presented in sections 3 and 4 in terms of biases with the observational estimates. We choose not to show results for the regional processes as these are generally for fewer models and only provide one sample of important features of North American climate. Other metrics, such as the RMSE, could have been used, but the bias values were available for all continental analyses. Model performance is shown by two methods: the first is the normalized bias, calculated as the difference of the absolute model bias from the lowest absolute bias value, divided by the range in absolute bias values across all models. A value of 0.0 indicates the lowest absolute bias and a value of 1.0 indicates the highest value. The values reflect the distribution of values across models such that outlier models are still identified. The second method is the rank of the sorted absolute bias values, which is uniformly distributed.

The first thing to note is the difference in spread between the two panels in Figure 19 for individual metrics, which is a reflection of the two types of distribution of values (model ensemble dependent and uniform). The normalized metrics highlight outlier models that perform much better than the rest of the models (e.g. BCC-CSM1.1 for runoff ratios (Q/P)) or much worse (e.g. INMCM4 for DJF precipitation or CanEMS2 for JJA temperature). Another example is persistent precipitation events (P persist US), for which there is a cluster of eight models that do equally poorly compared to the rest of the models. No single model stands out as being better or worse for multiple metrics. Some models do relatively well for the same variable and a single/both season across all regions, such as HadGEM2-ES and MIROC5 for precipitation.

The rankings in the lower panel are more clustered across analyses, such as for the Hadley Center models for DJF temperature, although the actual biases are generally not that different to the other models. The MRI-CGCM3 is consistently ranked low for runoff ratios in all regions and for the number of summer/frost days and growing season length (and in terms of the normalized metrics). The INMCM4 model is consistently ranked low for precipitation in both seasons and DJF temperature, although again its normalized values are generally not very different from the other models. It is tempting to provide an overall ranking or weighted metric across all analyses for each model, but there is no obivious way of doing this for a diverse set of metrics, although this has been attempted in other studies (e.g. Reichler and Kim, 2008). Nevertheless, it is useful to identify those models that are ranked highly for multiple metrics. For example, the following models are ranked in the top 5 for at least 12 metrics (approximately one third of the total number of metrics): MPI-ESM-LR (16 metrics), GISS-E2-R (15), CCSM4 (14), CSIRO-Mk3.6.0 (14), BCC-CSM1-1 (12); with the bottom two models: GFDL-CM3 (6) and INMCM4 (4).

6.2. Changes in Performance between CMIP3 and CMIP5 for Basic Climate Variables

A key question is whether the CMIP5 results have improved since CMIP3, and why. As mentioned in the introduction, the CMIP5 models generally have higher horizontal resolution and have improved parameterizations and additional process representations since CMIP3. Several of the analyses presented here indicate improved results since CMIP3, for example for the North American monsoon, by comparison with earlier studies. Here we show a direct comparison of CMIP5 with CMIP3 results for basic climate variables in Figure 20, which shows RMSE values for CMIP5 and CMIP3 models for seasonal precipitation and surface air temperature over North America and SSTs over the surrounding oceans. 14 of the 17 core CMIP5 models have an equivalent CMIP3 model, that is the same model (HadCM3), a newer version, or an earlier related version, and so a direct comparison of any improvements since CMIP5 is feasible.

Overall, the MME mean performance has improved slightly in CMIP5 for nearly all variables. For example, there is a reduction in the MME mean RMSE for summer precipitation (0.90 mm day^{-1} for CMIP3, 0.86 mm day^{-1} for CMIP5), and for winter SSTs (1.72 ^{o}C to 1.55 ^{o}C). The largest percentage reduction in RSME for the MME mean is for summer temperatures (11.8% reduction in RMSE). The spread in model performance (as quantified by the standard deviation) has remained about the same for precipitation, increased for temperature and decreased for SSTs. The increase in spread for temperature is due to both increases and decreases in model performance relative to the CMIP3 models. Several models have improved considerably and across nearly all variables and seasons, such as the CCSM4, INMCM4, IPSL-CM5A-LR, and MIROC5. Reductions in performance for individual models are less prevalent across variables, but are large for CSIRO-Mk3.6.0, HadCM3, and MRI-CGCM3 for SSTs in both seasons. The CanESM2 has worse performance than its CMIP3 equivalent (CGCM3.1) for all variables, although it is unclear how the two models are related. Interestingly the HadCM3 model, which is used for both the CMIP3 and CMIP5 simulations, appears to have degraded in performance for SSTs.

6.3. Summary and Conclusions

We have evaluated the CMIP5 multi-model ensemble for its depiction of North American continental and regional climatology, with a focus on a core set of models. Overall, the multi-model ensemble does reasonably well in representing the main features of basic surface climate over North America and the adjoining seas. Regional performance for basic climate variables is highly variable across models, however, and this can bias the assessment of the ensemble because of outlier models and therefore the median value may be a better representation of the central tendency of model performance (Liepert and Previdi, 2012). No particular model stands out as performing better than others across all analyses, although some models perform much better for sets of metrics, mainly for the same variable across different regions. Higher resolution models tend to do better at some aspects than others, especially for the regional features as expected, but not universally so and not for basic climate variables.

There are systematic biases in precipitation with overestimation for more humid and cooler regions and underestimation for drier regions. Biases in precipitation filter down to biases in the surface hydrology, although this is also related to the representation of the land surface in many models, with implications for assessment of water resources and hydrological extremes. The poor performance in representing observed seasonal persistence in precipitation and soil moisture is a reflection of this. As many of the errors are systematic across models, there is potential for diagnosing these further based on a multi-model analysis.

The models have a harder time representing extreme values, such as those based on temperature and precipitation. The biases in temperature means and extremes may be related to those in land hydrology that affects the surface energy balance and therefore can impact how much energy goes into heating the near surface air during dry periods and in drier regions. Biases in precipitation and its extremes are likely related to differences in large-scale circulation and SST patterns, as well as problems in representing regional climate features. Hints of this are shown in the some of the analyses presented here, such as the errors in regional moisture divergence over North America, but linkages between other regional climate features and terrestrial precipitation biases are not apparent, such as for western Atlantic winter cyclones, and further investigation is required to diagnose these. Part 2 of this paper (Sheffield et al., 2013) indicates that most models have trouble representing teleconnections between modes of climate variability (such as ENSO) and continental surface climate variables, and this may also reflect the representation of mean climate.

Overall, the performance of the CMIP5 models in representing observed climate features has not improved dramatically compared to CMIP3, at least for the set of models and climate features analyzed here. There are some models that have improved for certain features (e.g. the timing of the NAM), but others that have become worse (e.g. continental seasonal surface climate).

The results of this paper have implications for the robustness of future projections of climate and its associated impacts. Part three of this paper (Maloney et al., 2013) evaluates the CMIP5 models for N. America in terms of the future projections for the same set of climate features as evaluated for the 20th century in this first part and the second part of the paper (Sheffield et al., 2013). Whilst model historical performance is not sufficient for credible projections, the depiction of at least large scale climate features is necessary. Overall, the models do well in replicating the broad scale climate of N. America and some regional features, but biases in some aspects are of the same magnitude as the projected changes (Maloney et al., 2013). For example, the low bias in daily maximum temperature over the southern US in some models is similar to the future projected changes. Furthermore, the uncertainty in the future projections across models can also be of the same magnitude the model spread for the historic period.

Acknowledgements. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The authors acknowledge the support of NOAA Climate Program Office Modeling, Analysis, Predictions and Projections (MAPP) Program as part of the CMIP5 Task Force.

References

Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, and P. Arkin, 2003: The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *J. Hydrometeor.*, **4**,1147–1167.

Arora, V. K., J. F. Scinocca, G. J. Boer, J. R. Christian, K. L. Denman, G. M. Flato, V. V. Kharin, W. G. Lee, and W. J. Merryfield, 2011: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, *Geophys. Res. Lett.*, **38**, L05805, doi:10.1029/2010GL046270.

Bao, Q., and co-authors, 2012: The Flexible Global Ocean-Atmosphere-Land System model Version: FGOALS-s2. *Adv. Atmos. Sci*., submitted.

Bi, D., M. Dix, S. Marsland, T. Hirst, S. O'Farrell and coauthors, 2012: ACCESS: The Australian Coupled Climate Model for IPCC AR5 and CMIP5. AMOS conference, 2012, Sydney, Australia (available online at https://wiki.csiro.au/confluence/display/ACCESS/ACCESS+Publications)

Blackadar, A. K., 1957: Boundary layer wind maxima and their significance for their growth of nocturnal inversions. *Bull. Amer. Meteor. Soc.*, **38**, 283–290.

Bonner, W. D., 1968: Climatology of the low level jet. *Mon. Wea. Rev.*, **96**, 833-850.

Byerle, L. A., and J. Paegle, 2003: Modulation of the Great Plains low-level jet and moisture transports by orography and large scale circulations. *J. Geophys. Res.*, **108**, 8611, doi:10.1029/2002JD003005.

Caesar, J., L. Alexander, and R. Vose, 2006: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *J. Geophys. Res.*, **111**, D05101, doi:10.1029/2005JD006280

Chylek, P., J. Li, M. K. Dubey, M. Wang, and G. Lesins, 2011: Observed and model simulated 20th Century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmospheric Chemistry and Physics Discussions*, **11** (8), 22,893–22,907, doi: 10.5194/acpd-11-22893-2011

Colle, B. A., Z. Zhang, K. Lombardo, P. Liu, E. Chang, M. Zhang, and S. Hameed, 2013: Historical and future predictions of eastern North America and western Atlantic extratropical cyclones in CMIP5 during the cool season. *J. Climate*, submitted.

Collins, M., S. F. B Tett, and C. Cooper, 2001: The internal climate variability of HadCM3, a version of the Hadley Centre Coupled Model without flux adjustments. *Climate Dynamics*, **17** (1), 61-81.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S. D. Woodruff, and S. J. Worley, 2011: The Twentieth Century Reanalysis Project. *Quarterly J. Roy. Meteorol. Soc.*, **137**, 1-28. DOI: 10.1002/qj.776.

Donner, L. J., with 28 co-authors, 2011: The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL Global Coupled Model CM3. *J. Climate*, **24**, doi:10.1175/2011JCLI3955.1.

Dufresne, J-L., and 58 co-authors, N, 2012: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, *Clim. Dyn.*, submitted.

Fetterer, F., K. Knowles, W. Meier, and M. Savoie, 2002, updated 2009. /Sea Ice Index/. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.

Francis, J. A., and S. J. Vavrus, 2012: Evidence linking Arctic amplification to extreme weather in mid-latitudes. *Geophys. Res. Lett.*, 39, L06801, doi:10.1029/2012GL051000.

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, and T. Peterson, 2002: Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Res.*, **19**, 193–212.

Geil, K. L., Y. L. Serra, and X. Zeng, 2013: Assessment of CMIP5 model simulation of the North American Monsoon System. *J. Climate*, sumitted.

Gent, Peter R., and Coauthors, 2011: The Community Climate System Model Version 4. *J. Climate*, **24**, 4973–4991. doi: http://dx.doi.org/10.1175/2011JCLI4083.1

Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dynamics*, **16** (2-3), 169–182, doi:10.1007/PL00013733.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972

Hall, A., Qu, X., and Neelin, J. D., 2008: Improving predictions of summer climate change in the United States, *Geophys. Res. Lett.*, 35, L01702, DOI:10.1029/2007GL032012

Hazeleger, W., and 31 co-authors, 2010: EC-Earth: A seamless Earth system prediction approach in action. *Bull. Amer. Meteor. Soc.*, **91**, 1357-1363, doi: 10.1175/2010BAMS2877.1

Helfand, H. M., and S. D. Schubert, 1995: Climatology of the simulated Great Plains low-level jet and its contribution to the continental moisture budget of the United States. *J. Climate*, **8**, 784–806.

Higgins, R. W., J. E. Janowiak and Y.-P. Yao, 1996: A gridded hourly precipitation data base for the United States (1963-1993). NCEP/Climate Prediction Center Atlas No. 1, U. S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586.

Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123,** 3458–3465.

Holton, J. R., 1967: The diurnal boundary layer wind oscillation above sloping terrain. *Tellus*, **19**, 199–205.

Huffman, G. J., R. F. Adler, D. T. Bolvin, G. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F. Stocker, D. B. Wolff, 2007: The TRMM Multi–satellite Precipitation Analysis: Quasi–Global, Multi–Year, Combined–Sensor Precipitation Estimates at Fine Scale. *J. Hydrometeor.*, **8** (1), 38–55.

Jones, B. M., C. D. Arp, M. T. Jorgenson, K. M. Hinkel, J. A. Schmutz, and P. L. Flint , 2009: Increase in the rate and uniformity of coastline erosion in Arctic Alaska. *Geophys. Res. Lett.*, **36**, L03503, doi:10.1029/2008GL036205

Jones, C. D., and others, 2011: The HadGEM2-ES implementation of CMIP5 centennial simulations, *Geosci. Model Dev.*, **4,** 543-570, doi:10.5194/gmd-4-543-2011.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Woollen J., Zhu Y., Chelliah M., Ebisuzaki W., Higgins W., Janowiak J., Mo K.C., Ropelewski C., Wang J., Leetman A., Reynolds R., Jenne R., Joseph D., 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.*, **77**, 437–471.

Kanamitsu M., W. Ebisuzaki, J. S.Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.,* **83,** 1631-1643.

Kim, D., A. H. Sobel, A. D. Del Genio, Y. Chen, S. Camargo, M.-S. Yao, M. Kelley, and L. Nazarenko, 2012: The tropical subseasonal variability simulated in the NASA GISS general circulation model, *J. Clim.*, in press.

Kwok, R., and G. F. Cunningham, 2008: ICESat over Arctic sea ice: Estimation of snow depth and ice thickness. *J. Geophys. Res.*, **113**, C08010, doi:10.1029/2008JC004753.

Legates D. R., and G. J. McCabe, 1999: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35: 233-241.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for GCMs. *J. Geophys. Res.*, **99**, 14,415-14,428.

Liang, X.-Z., J. Zhu, K. E. Kunkel, M. Ting, and J. X. L. Wang, 2008: Do CGCMs simulate the North American monsoon precipitation seasonal-interannual variations. *J. Climate*, **21,** 3755-3775.

Liepert, B. G., and M. Previdi, 2012: Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models. *Environ. Res. Lett.*, 7 014006 doi:10.1088/1748-9326/7/1/014006

Maloney, E. D., S. J. Camargo, E. Chang, B. Colle, R. Fu, K. L. Geilw, Q. Hu, X. Jiang, N. Johnson, K. B. Karnauskas, J. Kinter, B. Kirtman, S. Kumar, B. Langenbrunner, K. Lombardo, L. Long, A. Mariotti, J. E. Meyerson, K. Mo, J. D. Neelin, Z. Pan, R. Seager, Y. Serraw, A. Seth, J. Sheffield, J. Thibeault, S.-P. Xie, C. Wang, B. Wyman, and M. Zhao, 2011: North American Climate in CMIP5 Experiments: Part III: Assessment of 21st Century Projections. *J. Climate*, submitted.

Mars, J. C., and D. W. Houseknecht, 2007: Quantitative remote sensing study indicates doubling of coastal erosion rate in past 50 yr along a segment of the Arctic coast of Alaska. *Geology*, **35** (7), 583-586. doi: 10.1130/G23672A.1

McCabe, G. J., T. R. Ault, B. I. Cook, J. L. Betancourt, and M. D. Schwartz, 2012: Influences of the El Nino Southern Oscillation and the Pacific Decadal Oscillation on the times of the North American spring. *Int. J. Climatol.*, 32: 2301-2310.

Mitchell, M. J., R. W. Arritt, and K. Labas, 1995: A climatology of the warm season Great Plains low-level jet using wind profiler observations. *Wea. Forecasting*, **10**, 576–591.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, **25**, 693–712.

Overland, J. E., 2011: Potential Arctic change through climate amplification processes. *Oceanography,* **24** (3), 176-185. http://dx.doi.org/10.5670/oceanog.2011.70

Rasmusson, E.M., 1967: Atmospheric water vapor transport and the water balance of the North America. Part I: Characteristics of the water vapor flux field. *Mon. Wea. Rev.,* **95**, 403-426.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Globally complete analyses of sea surface temperature, sea ice and night marine air temperature, 1871-2000. *J. Geophys. Res.*, **108** (4407).

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, 89, 303-311

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057. doi: http://dx.doi.org/10.1175/2010BAMS3001.1

Sakamoto, T. T., Y. Komuro, T. Nishimura, M. Ishii, H. Tatebe, H. Shiogama, A. Hasegawa, T. Toyoda, M. Mori, T. Suzuki, Y. Imada, T. Nozawa, K. Takata, T. Mochizuki, K. Ogochi, S. Emori, H. Hasumi and M. Kimoto, 2012: MIROC4h - a new high-resolution atmosphere-ocean coupled general circulation model. *J. Meteor. Soc. Japan*, **90** (3), 325-359.

Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling, *J. Climate*, **19** (13), 3088-3111.

Sheffield, J. and E. F. Wood, 2007: Characteristics of global and regional drought, 1950-2000: Analysis of soil moisture data from off-line simulation of the terrestrial hydrologic cycle. *J. Geophys. Res.*, **112** (D17), doi:10.1029/2006JD008288656

Sheffield, J., S. J. Camargo, B. Colle, Q. Hu, X. Jiang, N. Johnson, S. Kumar, K. Lombardo, B. Langenbrunner, E. Maloney, J. E. Meyerson, J. D. Neelin, Y. L. Serra, D.-Z. Sun, C. Wang, S.-P. Xie, J.-Y. Yu, T. Zhang, 2012: North American Climate in CMIP5 Experiments: Part II: Evaluation of 20^{th} Century Intra-Seasonal to Decadal Variability, *J. Climate*, submitted.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze, 2007: Arctic sea ice decline: Faster than forecast. *Geophys. Res. Lett.*, 34, L09501, doi:10.1029/2007GL029703

Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier, 2012: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations, *Geophys. Res. Lett.*, doi: 10.1029/2012GL052676R, in press.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Universidad Nacional Autónoma de México (UNAM), 2007: Gridded precipitation and temperature analysis from the Centro de Ciencias de la Atmósfera, Mexico; available from the International Research Institute for Climate and Society (http://ingrid.ldeo.columbia.edu/SOURCES/UNAM/gridded/monthly/v0705/).

Veres, M. C., and Q. Hu, 2012: AMO-forced regional processes affecting summertime precipitation variations in the central United States. *J. Climate*, in press.

Voldoire, A., and others, 2012: The CNRM-CM5.1 global climate model: Description and basic evaluation, *Clim. Dyn.*, doi:10.1007/s00382-011-1259-y, in press.

Volodin, E. M., N. A. Diansky, and A. V. Gusev, 2010: Simulating Present-Day Climate with the INMCM4.0 Coupled Model of the Atmospheric and Oceanic General Circulations. Izvestia, *Atmospheric and Oceanic Physics*, **46**, 414-431

Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. Eischeid, 1992: The Global Historical Climatology Network: long-term monthly temperature, precipitation, sea level pressure, and station pressure data. ORNL/CDIAC-53, NDP-041. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Wang, C., and D. B. Enfield, 2001: The tropical Western Hemisphere warm pool. *Geophys. Res. Lett.*, **28**, 1635-1638.

Wang, M. and J. E. Overland (2009), A sea ice free summer Arctic within 30 years?, *Geophys. Res. Lett.*, **36**, L07502, doi:10.1029/2009GL037820.

Wang, A., T. J. Bohn, S. P. Mahanama, R. D. Koster, and D. P. Lettenmaier, 2009: Multimodel ensemble reconstruction of drought over the continental United States. *J. Climate*, **22**, 2694–2712.

Watanabe, M., and Coauthors, 2010: Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. *J. Climate*, **23**, 6312–6335.

Wexler, H., 1961: A boundary layer interpretation of the low level jet. *Tellus*, **13**, 368–378.

Xie, P., and P.A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539 - 2558.

Xie, P., M. Chen, and W. Shi, 2010: CPC unified gauge-based analysis of global daily precipitation. Preprints, *24th Conf. on Hydrology,* Atlanta, GA, Amer. Meteor. Soc., 2.3A

Xin X., Wu T., Zhang J., 2012: Introductions to the CMIP 5 simulations conducted by the BCC climate system model (in Chinese). *Advances in Climate Change Research.* submitted.

Yukimoto, S., et al., 2012: A new global climate model of the Meteorological Research Institute: MRI-CGCM3—Model description and basic performance, *J. Meteorol. Soc. Jpn.*, **90a**, 23–64.

Zanchettin, D., A. Rubino, D. Matei, O. Bothe, and J. H. Jungclaus, 2012: Multidecadal-to-centennial SST variability in the MPI-ESM simulation ensemble for the last millennium. *Clim. Dyn.*, **39**, 419-444 doi:10.1007/s00382-012-1361-9.

Zhang, Z. S., Nisancioglu, K., Bentsen, M., Tjiputra, J., Bethke, I., Yan, Q., Risebrobakken, B., Andersson, C., and Jansen, E., 2012: Pre-industrial and mid-Pliocene simulations with NorESM-L. *Geosci. Model Dev.*, **5**, 523-533, doi:10.5194/gmd-5-523-2012.

Figure Captions

Figure 1. Precipitation climatology for (left) December-February and (right) June-August (1979-2005). a) GPCP estimate of observed precipitation for DJF. b) MME mean over the 18 models for DJF; for models with multiple runs, all runs are averaged before inclusion in the multi-model ensemble. c) Comparison of individual models to observations using the 3 mm day^{-1} contour as an index of the major precipitation features: half the models are shown in each of sub-panel I and II with the legend giving the color-coding for the models in each. Shading shows the regions where GPCP exceeds 3 mm day^{-1}; a model with no error would have its contour fall exactly along the edge of the shaded region. d)-f) As in a)-c), respectively, except for JJA.

Figure 2. Surface air temperature climatology for (top) December-February and (bottom) June-August (1979-2005). a) MME mean (over the 17 core models plus FGOALS-s2) for DJF. b) NCEP-DOE Reanalysis 2 estimate of observed surface air temperature climatology for DJF. c) As in b) but for CRU. d) Standard deviation of surface air temperature among the 18 model DJF climatological values at each point. e) Difference between the MME mean climatology in a) and the NCEP-DOE Reanalysis 2. f) As in e) but for CRU. g)-l) As in a)-f) but for JJA.

Figure 3. Climatological sea surface temperature and precipitation in observations from HadISSTv1.1 and GPCPv2.2 data sets, and historical simulations from 17 CMIP5 models for 1979-2004. (a) Observations, (b) MME mean, and (c) MME mean minus observations, for winter-to-spring (December to May). (d-f) as in (a-c) but for summer-to-fall (April to November). Temperatures are shaded blue/red for values equal or lower/larger than 23/24°C; the thick black line highlights the 28.5°C isotherm as indicator of the Western Hemisphere Warm Pool. Precipitation is shaded green for values equal or larger than 2 mm day^{-1}. Contour intervals are 1°C and 1 mm day^{-1} for the mean values and 0.2°C and 2 mm day^{-1} for the differences. SST/precipitation fields have been regridded to a common 5°×2.5°/2.5°×2.5° grid.

Figure 4. Vertically integrated moisture transport (vectors) and its divergence (contours) for the 20CR reanalysis (a,g) and five CMIP5 models for mean JJA (b-f) and DJF (h-l) for 1981-2000. Vertically integrated moisture transport is computed to 500 hPa using 6-hourly data from the 20CR and one realization each from the historical experiments for CanESM2, CCSM4, CNRM-CM5, GFDL-ESM2M, and MIROC5 models.

Figure 5. Mean seasonal cycle (1979-2005) of North American regional land water budget components for 12 CMIP5 models (CanESM2, CSIRO-Mk3-6-0, GFDL-ESM2G, GISS-E2-H, GISS_E2-R, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC-ESM, MIROC-ESM-CHEM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M) compared to the average of the two off-line LSM simulations (VIC and GLDAS2 Noah). Regions are Western North America (WNA), Central North America (CNA), Eastern North America (ENA), Alaska and Western Canada (ALA), Northeast Canada (NEC), and Central America (CAM) as modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3.

Figure 6. Mean annual runoff (mm/year) (top) and runoff (Q/P) ratio (bottom) for 1979-2004 from observations (Q from VIC and GLDAS2 Noah, and P from GPCP) and the multi-model average from 15 CMIP5 climate models (BCC-CSM1-1, CanESM2, CCSM4, CNRM-CM5, CSIRO-Mk3-6-0, GFDL-CM3, GFDL-ESM2M, GISS-E2-R, HadCM3, INMCM4, MIROC5, MIROC-ESM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M).

Figure 7. Comparison of biophysical indicators between observations and the CMIP5 ensemble. Biophysical indicators are (top row) number of summer days, (middle row) number of frost days, and (bottom row) growing season length averaged over 1979-2005. Left column shows the observations (left column) from the HadGHCND dataset; middle column is the multi-model ensemble mean of the 17 core models, and the right column is their difference (MME – obs.). The frequencies are calculated on the model grid and then interpolated to 2.0 degree resolution for comparison with the observational estimates.

Figure 8. The frequency of occurrence of persistent extreme precipitation events defined by SPI6 averaged over positive and negative events for (a) observed precipitation based on the CPC and UW datasets, (b) BCC-CSM1-1, (c) CanESM2, (d) CCSM4, (e) CNRM-CM5.1, (f) CSIRO-Mk3.6.0, (g) GFDL-CM3, (h) GISS-E2-R, (i) HadCM3, (j) IPSL-CM5A-LR, (k) MIROC5, (l) MIROC-ESM, (m) MPI-ESM-LR, (n) MRI-CGCM3 and (o) NorESM1-M. The HadCM3 and HadGEM2-ES results are similarly weak and so the former are shown only. Each data set is treated as one member of the ensemble.

Figure 9. Same as Figure 8 but for persistent soil moisture events. Estimates of observed soil moisture are taken from the multi-model NLDAS-UW dataset.

Figure 10. (a) Cyclone density for the CFSR analysis showing the number of cyclones per cool season (November to March) per 50,000 km^2 for 1979-2004. (b) Same as (a) except for the mean (shaded) and spread (contoured every 0.3) of 15 CMIP5 models ordered from higher to lower spatial resolution: CanESM2, EC-EARTH, MRI-CGCM3, CNRM-CM5, MIRCO5, HadGEM2-ES, HadGEM2-CC, INMCM4, IPSL-CM5A-MR, MPI-ESM-LR, NorESM1-M, GFDL-ESM2M, IPSL-CM5A-LR, BCC-CSM1, MIROC-ESM-CHEM. Same as (a) except for the (c) MPI-ESM-LR, (d) GFDL-ESM2M, (e) HadGEM2-CC, and (f) CCSM4 models.

Figure 11. Number of cyclone central pressures at their maximum intensity (minimum pressure) for the 1979-2004 cool seasons within the dashed box region in Fig. 10 for a 10 hPa range centered every 10 hPa showing the CFSR (bold blue), (b) CMIP5 MME mean (bold red), and individual CMIP5 models.

Figure 12. (a) CPC merged precipitation analysis at 2.5 deg resolution showing cool seasonal average precipitation (shaded every 75 mm) for the 1979-2004 cool seasons (November – March). (b) Same as (a) except for the CPC Unified precipitation at 0.5 deg resolution. (c) Same as (a) except for the mean of 14 of the 17 CMIP5 members listed in (d) and spread (in mm). (d) Number of days that the daily average precipitation (in mm/day) for the land areas in the black box in (b) occurred within each amount bin for select CMIP5 members, CMIP5 mean, and the CPC Unified.

Figure 13. Comparison of precipitation and temperature extremes for southern US regions between the CMIP5 models and CPC and GHCN observations, respectively. (left column) Taylor diagram of the spatial pattern of annual number of days when precipitation > 10mm day^{-1} over the southwest (SW), south central (SC) and southeastern (SE) US. (right column) Taylor diagram of the spatial pattern of annual number of days when Tmax > 32^{o}C (90F) for the three regions. The standard deviations have been normalized relative to the observed values. (A: CanESM2, B: CCSM4, C: GFDL-CM3, D: GFDL-ESM2G, E: GFDL-ESM2M, F: GISS-E2-R, G: HadCM3, H: HadGEM2-CC, I: HadGEM2-ES, J: IPSL-CM5A-LR, K: MIROC4h, L: MIROC5, M: MPI-ESM-LR, N: MRI-CGCM3). Observations are from the CPC dataset. SW is defined as the contiguous US south of 40^{o}N between 125^{o}W and 110^{o}W; SC is the contiguous US south of 40^{o}N between 110^{o}W and 90^{o}W; SE is the contiguous US south of 40^{o}N between 90^{o}W and 70^{o}W.

Figure 14. Average monthly precipitation for 1979-2005 shown by latitude in the North American monsoon region (longitudes 102.5 to 115W) from the CMAP observational estimate (a), the MME mean for the 17 core CMIP5 models (b) and their difference (c), all in units of mm day^{-1}.

Figure 15. Annual cycle in rainfall for the NAM region for the historical (1979-2005) period of 21 CMIP5 models compared to the P-NOAA AND CMAP observational datasets for (a) small (phase error = 0), (b) moderate (phase error = 1), (c) large (phase error = 2-4) phase errors, and (d) all models.

Figure 16. (a)-(c) Averaged summer 925hPa wind during 1971-2000 for NCEP-NCAR reanalysis, eight-model CMIP5 ensemble mean for the same period, and the reanalysis minus MME mean, respectively. (d)-(f) Lower troposphere mean vertical profile of meridional wind averaged over 95°-100°W for the reanalysis, MME mean, and the reanalysis minus MME mean, respectively. (g)-(i) Seasonal cycle of the 925hPa meridional wind averaged over 27.5º-32.5ºN for the reanalysis, MME mean, and the reanalysis minus MME mean. All units are m s^{-1}. Shading indicates wind speeds greater than 3.0 m s^{-1} in the figures of the first and second columns and wind speeds greater than 1.0m s^{-1} in the figures of the third column.

Figure 17. September and March sea ice extent from 26 CMIP5 models compared to observations from the NSIDC from 1953 to 2005. For each model, the boxes represent inter-quartile ranges (25th to 75th percentiles). Median (50th percentile) extents are shown by the thick horizontal bar in each box. The width of each box corresponds to the number of ensemble members for that model. Whiskers (vertical lines and thin horizontal bars) represent the 10th and 90th percentiles. Mean monthly extents are shown as diamonds. Corresponding mean, minimum and maximum observed extends are shown as red and green lines, respectively.

Figure 18. March (left) and September (right) ice thickness (m) for 26 CMIP5 models averaged over 1993-2005 versus satellite and airborne observations for ERS1/2 (1993-2001), ICESat (2003-2009) and IceBridge (2009-2012).

Figure 19. Comparison of CMIP5 models across a set of continental performance metrics based on bias values given in Tables 3-8. (top) Biases normalized relative to the range of bias values across models, with lower values indicating lower bias. (bottom) Models ranked according to bias values, with 1 indicating the model with the lowest bias and 17 the model with the highest bias. Results for models without available data are indicated in white. The bias metrics shown (in order from left to right) are for regional precipitation (P) for DJF and JJA, regional temperature (T) for DJF and JJA, annual SSTs for surrouding oceans (see Figure 3), annual runoff ratios (Q/P), the annual number of summer days (SuDays), frost days (FrDays) and growing season length (GSL), and east-west gradient in the number of persistent precipitation (P Persist) and soil moisture (SM Persist) events.

Figure 20. Comparison of CMIP5 and CMIP3 model performance for seasonal (DJF and JJA) precipitation (P), surface air temperature (T) and SST. Results are shown as RMSE values calculated for 1971-1999 relative to the GPCP, CRU and HadISST observational datasets. Precipitation and temperature RMSE values are calculated over North America (130-60W, 0-60N) and SST RMSE values are calculated over neighboring oceans (170-35W, 10S-40N). The core set of CMIP5 models and their equivalent CMIP3 models where available (otherwise indicated by N/A) are shown. The MME mean values are also shown.

Table 1. CMIP5 models evaluated and their attributes. The core models are highlighted with an asterix.

Model	Center	Atmospheric Horizontal Resolution (lon. x lat.)	Number of model levels	Reference
ACCESS1-0	Commonwealth Scientific and Industrial Research Organization/Bureau of Meteorology, Australia	1.875 x 1.25	38	Bi et al. (2012)
BCC-CSM1.1*	Beijing Climate Center, China Meteorological Administration, China	2.8 x 2.8	26	Xin et al. (2012)
CanCM4	Canadian Centre for Climate Modelling and Analysis, Canada	2.8 x 2.8	35	Chylek et al. (2011)
CanESM2*	Canadian Center for Climate Modeling and Analysis, Canada	2.8 x 2.8	35	Arora et al. (2011)
CCSM4*	National Center for Atmospheric Research, USA	1.25 x 0.94	26	Gent et al. (2011)
CESM1-CAM5-1-FV2	Community Earth System Model Contributors (NSF-DOE-NCAR)	1.4 x 1.4	26	Gent et al. (2011)
CNRM-CM5.1*	National Centre for Meteorological Research, France	1.4 x 1.4	31	Voldoire et al. (2011)
CSIRO-MK3.6*	Commonwealth Scientific and Industrial Research Organization/Queensland Climate Change Centre of Excellence, AUS	1.8 x 1.8	18	Rotstayn et al. (2010)
EC-EARTH	EC-EARTH consortium	1.125 x 1.12	62	Hazeleger et al. (2010)
FGOALS-S2.0	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	2.8 x 1.6	26	Bao et al. (2012)
GFDL-CM3*	NOAA Geophysical Fluid Dynamics Laboratory, USA	2.5 x 2.0	48	Donner et al. (2011)
GFDL-ESM2G/M*	NOAA Geophysical Fluid Dynamics Laboratory, USA	2.5 x 2.0	48	Donner et al. (2011)
GISS-E2-H/R*	NASA Goddard Institute for Space Studies, USA	2.5 x 2.0	40	Kim et al. (2012)
HadCM3*	Met Office Hadley Centre, UK	3.75 x 2.5	19	Collins et al. (2001)

HADGEM2-CC (Chemistry coupled)	Met Office Hadley Centre, UK	1.875 x 1.25	60	Jones et al. (2011)
HadGEM2-ES*	Met Office Hadley Centre, UK	1.875 x 1.25	60	Jones et al. (2011)
INMCM4*	Institute for Numerical Mathematics, Russia	2 x 1.5	21	Volodin et al. (2010)
IPSL-CM5A-LR*	Institut Pierre Simon Laplace, France	3.75 x 1.8	39	Dufresne et al. (2012)
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	2.5 x 1.25	39	Dufresne et al. (2012)
MIROC4h	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	0.56 x 0.56	56	Sakamoto et al. (2012)
MIROC5*	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	1.4 x 1.4	40	Watanabe et al. (2010)
MIROC-ESM*	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	2.8 x 2.8	80	Watanabe et al. (2010)
MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	2.8 x 2.8	80	Watanabe et al. (2010)
MPI-ESM-LR*	Max Planck Institute for Meteorology, Germany	1.9 x 1.9	47	Zanchettin et al. (2012)
MRI-CGCM3*	Meteorological Research Institute, Japan	1.1 x 1.1	48	Yukimoto et al. (2011)
NorESM1-M*	Norwegian Climate Center, Norway	2.5 x 1.9	26	Zhang et al. (2012)

Table 2. Observational and reanalysis datasets used in the evaluations

Dataset	Type	Spatial Domain	Temporal Domain	Reference
		Precipitation		
CMAP v2	Gauge/satellite	2.5 deg, global	Monthly/pentad, 1979-present	Xie and Arkin, 1997
GPCP v2.1	Gauge/satellite	1.0 deg, global	Monthly, 1979-2009	Adler et al., 2003
CRU TS3.1	Gauge	0.5 deg, global land	Monthly, 1901-2008	Mitchell and Jones (2005)
CMAP v2	Gauge/satellite	2.5 deg, global	Monthly/pentad, 1979-present	Xie and Arkin, 1997
CPC unified	Gauge	0.5 deg, US	Daily, 1948-2010	Xie et al., 2010
CPC-US-Mexico	Gauge	1.0 deg, US/Mexico	Daily, 1948-present	Higgins et al. (1996)
UW	Gauge	0.5 deg, US	Daily, 1916-2009	Maurer et al. (2002)
P-NOAA	Gauge	0.5 deg, North America	Monthly, 1895-2010	Cook and Vose, 2011*
		Temperature		
CRU TS3.1	Gauge	0.5 deg, global land	Monthly, 1901-2008	Mitchell and Jones (2005)
GHCN	Gauge	2.5 degree, global land	Daily, varies	Vose et al. (1992)
HadGHCND	Gauge	2.5x3.75 degree,	Daily, 1950-2000	Caesar et al. (2006)

		global land		
		Sea Surface Temperature and Sea Ice		
HadISSTv1.1	In-situ/satellite	1.0 deg, global oceans	Monthly, 1870-present	Rayner et al. (2003)
NSIDC Sea Ice Index	Satellite	Arctic Basin	Monthly, 1979-present	Fetterer et al., 2002
IceSAT	Satellite	25km, Arctic Basin	Monthly, 2003-2008	Kwok and Cunningham (2008)
		Land Surface Hydrology		
NLDAS-UW	Multiple LSMs	0.5 deg US	Daily, 1916-2009	Wang et al. (2009)
VIC	VIC LSM	1.0 deg, global land	3-hourly, 1948-2008	Sheffield et al. (2007)
GLDAS	Noah LSM	1.0 deg, global land	3-hourly, 1979-2008	Rodell et al. (2004)
		Reanalyses		
NCEP-NCAR	Model reanalysis	~1.9 deg, global	6-hourly, 1948-present	Kalnay et al. (1996)
NCEP-DOE	Model reanalysis	~1.9 deg, global	6-hourly, 1979-present	Kanamitsu et al. (2002)
CFSR	Model reanalysis	~0.3 deg, global	6-hourly, 1979-2010	Saha et al. (2010)
20CR	Model reanalysis	2.0 deg, global	6-hourly, 1871-present	Compo et al. (2011)

*P-NOAA dataset provided by Drs. Russ Vose and Ed Cook.

Table 3. DJF and JJA bias (% of observed mean) in CMIP5 continental and regional precipitation relative to the GPCP observations. North America (NA): 10N to 72N and 190E to 305E; contiguous US (conUS): 25N to 50N and 235W to 285W; the regions are Alaska (ALA), North East Canada (NEC), Eastern North America (ENA), Central North America (CNA), West North America (WNA), and Central America (CAM), as modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3.

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
	DJF							
BCC-CSM1-1	14.84	21.39	40.32	-1.40	-10.24	-16.38	47.51	97.71
CanESM2	-3.80	-5.53	1.24	9.81	-2.23	-13.22	5.26	-21.09
CCSM4	17.13	10.24	49.31	16.52	-7.87	-20.21	48.87	64.19
CNRM-CM5	0.44	-1.58	4.50	-0.05	-5.71	-25.48	27.34	1.55
CSIRO-Mk3-6-0	2.82	5.35	16.78	7.36	-11.15	-8.46	20.41	13.45
GFDL-CM3	18.92	27.41	19.57	7.98	4.46	-2.60	52.70	59.52
GFDL-ESM2M	10.09	14.70	4.50	11.77	0.03	-16.67	40.63	45.80
GISS-E2-R	23.68	25.18	32.55	15.27	1.45	-6.81	47.96	110.22
HadCM3	2.59	14.43	-9.81	0.84	10.82	-4.61	27.56	-25.27
HadGEM2-ES	1.38	4.26	-4.11	-9.20	6.66	-4.44	13.90	-4.44
INMCM4	40.75	25.34	75.24	56.06	14.36	-7.40	71.31	97.49
IPSL-CM5A-LR	16.64	12.38	53.18	18.24	4.84	-14.11	42.28	25.43
MIRCO5	7.51	7.94	14.83	11.28	-1.46	-13.67	29.22	28.43
MIROC-ESM	16.82	8.73	66.33	31.07	-2.38	-37.19	70.34	-12.75
MPI-ESM-LR	11.87	14.27	11.71	21.56	3.17	-6.02	34.18	18.59
MRI-CGCM3	24.97	34.43	27.12	-6.09	8.94	-6.16	70.00	85.93
NorESM1-M	2.17	-6.86	59.74	4.10	-28.15	-39.07	25.49	65.74
MME mean	12.28	12.48	27.24	11.48	-0.85	-14.26	39.70	38.27
	JJA							
BCC-CSM1-1	-14.95	-20.74	22.05	-8.61	-20.63	-30.17	-10.90	-16.91
CanESM2	-25.72	-39.23	40.26	-10.17	-22.26	-49.66	-25.63	-49.05
CCSM4	-7.04	3.47	23.70	0.92	7.67	-5.36	4.06	-39.06
CNRM-CM5	-0.13	-4.73	42.30	8.02	10.39	-29.25	31.69	-25.13
CSIRO-Mk3-6-0	-3.52	-29.14	54.57	4.81	-17.30	-38.92	-3.67	6.61
GFDL-CM3	7.65	15.95	34.39	20.85	-4.38	7.45	54.74	-31.10
GFDL-ESM2M	7.35	12.03	61.15	21.49	-5.20	2.66	49.86	-31.24
GISS-E2-R	20.81	32.35	14.93	11.93	22.82	24.24	66.36	-9.00
HadCM3	-3.44	1.16	26.83	8.19	0.57	0.73	19.82	-40.07
HadGEM2-ES	-1.41	-15.89	72.29	14.40	-0.04	-32.63	4.59	-16.90
INMCM4	4.00	-1.74	43.12	31.19	17.89	-19.67	44.51	-44.06
IPSL-CM5A-LR	-13.67	-1.62	16.88	3.27	9.85	-13.26	13.83	-63.80
MIRCO5	3.32	0.72	38.38	-3.87	10.09	-17.38	36.93	-13.17

MIROC-ESM	-4.37	4.37	47.01	6.18	1.62	-27.95	69.36	-56.12
MPI-ESM-LR	9.05	11.02	37.85	19.95	14.30	8.21	10.36	-14.15
MRI-CGCM3	10.98	16.85	23.16	9.44	6.35	0.48	61.05	-10.33
NorESM1-M	-10.01	11.48	12.31	-9.71	2.99	12.31	9.47	-51.51
MME mean	-1.24	-0.22	35.95	7.55	2.04	-12.25	25.67	-29.70

Table 4. As Table 3, but for temperature with bias in ^{o}C.

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
	DJF							
BCC-CSM1-1	-0.97	-1.98	1.95	-1.61	-1.29	-1.30	-1.88	-1.20
CanESM2	2.09	2.04	1.46	3.13	4.79	4.06	0.74	-0.84
CCSM4	0.01	-0.53	1.81	-0.30	1.03	-0.22	-0.37	-2.04
CNRM-CM5	-1.87	-1.95	-3.88	-0.99	-0.54	-0.55	-2.39	-2.54
CSIRO-Mk3-6-0	-1.61	-2.05	-2.31	-1.51	-1.04	-2.27	-1.31	-1.92
GFDL-CM3	1.37	-0.03	4.07	3.04	1.15	1.25	0.27	-2.31
GFDL-ESM2M	1.67	-0.84	5.05	5.75	1.35	0.75	-0.40	-3.31
GISS-E2-R	-0.33	-1.21	0.72	1.00	-0.56	-0.51	-1.13	-1.23
HadCM3	-2.88	-3.46	-3.72	-2.16	-1.51	-3.53	-3.48	-2.00
HadGEM2-ES	-3.81	-2.98	-7.00	-5.12	-1.61	-3.15	-3.47	-1.54
INMCM4	1.71	0.26	3.73	3.38	1.87	1.48	1.51	-3.71
IPSL-CM5A-LR	0.15	-1.14	4.16	-0.12	0.00	-0.64	-0.64	-2.27
MIRCO5	1.02	0.65	0.74	2.14	1.78	1.16	0.27	0.39
MIROC-ESM	3.29	1.74	4.22	6.74	3.90	3.00	1.35	0.73
MPI-ESM-LR	-0.08	0.30	-1.85	1.50	0.64	1.28	-0.93	-0.10
MRI-CGCM3	-0.37	-0.37	1.78	-2.63	-1.33	0.38	0.56	-2.26
NorESM1-M	-0.89	-1.60	2.12	-2.14	-0.34	-1.44	-1.48	-1.81
MME mean	-0.09	-0.77	0.77	0.59	0.49	-0.01	-0.75	-1.65
	JJA							
BCC-CSM1-1	-0.76	0.54	-3.24	-1.40	0.16	1.80	-0.78	-0.20
CanESM2	3.14	4.11	1.77	3.17	3.14	5.76	3.60	0.62
CCSM4	1.10	1.37	-0.13	1.55	1.21	2.04	1.78	-0.82
CNRM-CM5	0.41	-0.06	1.57	1.23	-0.72	1.16	0.08	-1.37
CSIRO-Mk3-6-0	1.30	2.97	-1.25	-0.11	1.38	5.18	1.45	1.46
GFDL-CM3	-1.96	-2.24	-1.47	-2.60	-1.37	-2.27	-2.04	-1.65
GFDL-ESM2M	-0.36	-0.59	-0.29	0.25	-0.34	-0.19	-0.75	-0.67
GISS-E2-R	-0.64	-2.09	1.71	0.79	-0.46	-1.74	-2.10	-1.25
HadCM3	-1.12	-0.74	-1.41	-1.73	-1.70	0.08	-1.30	-0.23
HadGEM2-ES	1.17	1.79	0.98	0.29	0.95	2.56	2.22	-1.36
INMCM4	-0.82	-2.07	0.78	1.23	-1.68	-1.02	-2.05	-2.52
IPSL-CM5A-LR	0.38	-1.22	2.60	2.45	0.35	0.01	-1.43	-1.40
MIRCO5	2.63	2.30	3.88	2.56	2.15	3.11	2.79	0.26
MIROC-ESM	2.76	1.89	2.91	4.09	3.03	3.31	1.67	2.20
MPI-ESM-LR	-0.84	-0.66	-1.31	-0.40	-0.81	-0.09	-1.55	0.02
MRI-CGCM3	-0.37	-1.77	2.10	1.23	-0.77	-0.97	-1.80	-2.04
NorESM1-M	-1.02	-0.51	-3.05	-0.90	-0.87	-0.64	0.03	-1.35

MME mean	0.29	0.18	0.36	0.69	0.21	1.06	-0.01	-0.61

Table 5. CMIP5 error statistics for annual average SSTs. Statistics are calculated over neighboring oceans to North America (170-35^{o}W, 10S-40^{o}N; domain displayed in Fig. 3) for average annual values for 1979-2004.

Model	Spatial Correlation	RMSE (^{o}C)	Bias (^{o}C)
BCC-CSM1.1	0.95	0.98	-0.15
CanESM2	0.97	0.87	-0.16
CCSM4	0.96	0.88	0.37
CNRM-CM5.1	0.96	0.94	-0.63
CSIRO-MK3.6	0.94	1.39	-1.80
GFDL-CM3	0.94	1.05	-0.68
GFDL-ESM2M	0.94	1.08	-0.46
GISS-E2-R	0.95	0.98	-0.15
HadCM3	0.94	1.53	-0.64
HadGEM2-ES	0.96	0.95	-0.96
INMCM4	0.94	1.06	-0.01
IPSL-CM5A-LR	0.93	1.32	-0.78
MIROC5	0.95	0.95	-0.65
MIROC-ESM	0.91	1.32	-0.60
MPI-ESM-LR	0.96	0.95	-0.49
MRI-CGCM3	0.95	1.25	-0.49
NorESM1-M	0.92	1.21	-0.78
MME mean	0.97	0.77	-0.54

Table 6. Spatial correlations between simulated and observed estimates of divergence for summer (JJA) and winter (DJF) seasons for the North American region. The CMIP5 model data were regridded to the 20CR grid for this calculation.

Model	Spatial Correlation
	Summer (JJA)
CanESM2	0.28
CCSM4	0.18
CNRM-CM5	0.39
GFDL-ESM2M	0.08
MIROC5	0.42
	Winter (DJF)
CanESM2	0.76
CCSM4	0.72
CNRM-CM5	0.75
GFDL-ESM2M	0.66
MIROC5	0.60

Table 7. Bias (model minus observations) in annual runoff ratio (total runoff / precipitation) averaged over 1979-2004 for the North American continent, the contiguous US and the six regions defined in Table 3.

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
BCC-CSM1-1	0.09	0.06	0.36	-0.10	-0.04	-0.03	0.11	0.21
CanESM2	0.40	0.26	0.69	0.56	0.35	0.20	0.43	0.12
CCSM4	0.36	0.19	0.75	0.51	0.31	0.10	0.37	0.15
CNRM-CM5	0.38	0.26	0.54	0.52	0.39	0.17	0.42	0.22
CSIRO-Mk3-6-0	0.30	0.19	0.53	0.37	0.28	0.16	0.26	0.26
GFDL-CM3	0.33	0.24	0.47	0.47	0.37	0.18	0.35	0.16
GFDL-ESM2M	0.32	0.25	0.36	0.38	0.36	0.16	0.36	0.30
GISS-E2-R	0.22	0.19	0.15	0.28	0.28	0.14	0.27	0.20
HadCM3	0.37	0.29	0.57	0.54	0.43	0.31	0.31	0.03
INMCM4	0.37	0.21	0.64	0.55	0.33	0.12	0.40	0.11
MIRCO5	0.33	0.21	0.58	0.49	0.40	0.17	0.29	0.23
MIROC-ESM	0.37	0.25	0.56	0.49	0.44	0.12	0.36	0.10
MPI-ESM-LR	0.31	0.21	0.55	0.46	0.33	0.18	0.30	0.11
MRI-CGCM3	0.42	0.32	0.64	0.55	0.46	0.28	0.40	0.28
NorESM1-M	0.35	0.18	0.77	0.51	0.30	0.11	0.34	0.15
MME mean	0.33	0.22	0.54	0.44	0.33	0.16	0.33	0.18

Table 8. Bias and spatial correlation between the HadGHCND observations and the CMIP5 ensemble for number of summer days, number of frost days and growing season length averaged over 1979-2005.

	Number of summer days		Number of frost days		Growing season length (days)	
Model	Bias (days)	Spatial correlation	Bias (days)	Spatial correlation	Bias (days)	Spatial correlation
BCC-CSM1-1	-14.0	0.95	-4.7	0.96	-7.8	0.91
CanESM2	17.1	0.96	-16.4	0.93	-12.1	0.90
CCSM4	0.0	0.88	-3.5	0.95	-9.0	0.92
CNRM-CM5	-7.4	0.92	12.6	0.92	-14.2	0.89
CSIRO-Mk3-6-0	-8.2	0.98	3.7	0.95	-4.3	0.90
GFDL-CM3	-39.5	0.93	0.6	0.97	-24.6	0.93
GFDL-ESM2M	33.0	0.92	-7.8	0.96	-5.6	0.92
GISS-E2-R	33.5	0.94	-12.8	0.96	7.4	0.96
HadCM3	-21.9	0.98	21.6	0.95	-38.2	0.88
HadGEM2-ES	-6.9	0.92	2.2	0.97	-14.9	0.95
INMCM4	-28.8	0.94	17.0	0.85	-76.1	0.53
IPSL-CM5A-LR	-39.3	0.85	4.1	0.97	-6.5	0.95
MIROC5	1.3	0.91	-15.1	0.98	33.4	0.96
MIROC-ESM	-5.7	0.90	-34.7	0.97	38.5	0.96
MRI-CGCM3	-34.8	0.87	-6.8	0.95	4.0	0.95
MPI-ESM-LR	-30.4	0.92	-12.5	0.94	5.5	0.93
NorESM1-M	-21.6	0.89	-3.7	0.96	-19.7	0.93
MME mean	-18.1	0.96	-2.8	0.97	-8.5	0.95

Table 9. Frequency of occurrence of persistent extreme precipitation and soil moisture events over the US for the CPC observations/NLDAS analysis and 15 CMIP5 models.

Model	SPI6	SM
Obs/Analysis	0.37	0.68
BCC-CSM1.1	0.00	0.05
CanESM2	0.29	0.32
CCSM4	0.35	0.14
CNRM-CM5.1	0.16	0.30
CSIRO-Mk3.6.0	0.02	0.70
GFDL-CM3	0.02	0.01
GISS-E2-R	0.04	0.47
HadCM3	0.00	0.32
HadGEM2-ES	0.00	0.37
IPSL-CM5A-LR	0.21	0.20
MIROC5	0.34	0.25
MIROC-ESM	0.26	0.27
MPI-ESM-LR	0.23	0.62
MRI-CGCM3	0.16	0.01
NorESM1-M	0.01	0.08

Table 10: Error statistics for the CMIP5 model precipitation over the northeastern US. The mean absolute error (mm per season), RMSE (mm day^{-1}), and mean bias (model/observed) for 14 CMIP5 models verified using the daily CPC-Unified precipitation within the black box in Fig. 10b.

Model	Mean Absolute Error (mm/season)	Root Mean Square Error (mm/day)	Mean Bias (model/obs)
MME mean	84.55	0.89	1.10
CanESM2	94.02	1.08	1.04
CCSM4	101.07	1.06	1.10
GFDL_ESM2M	101.32	1.15	1.13
GFDL_CM3	103.53	1.14	1.14
BCC-CSM1-1	104.62	1.16	1.08
CNRM_CM5	104.99	1.12	1.16
HadGEM_ES	105.41	1.18	1.16
MIROC_ESM	111.21	1.25	0.92
NorESM1_M	112.70	1.23	1.08
CSIRO_Mk_3_6_0	114.49	1.46	1.03
IPSL_CM5A_LR	115.96	1.27	1.03
MIROC5	118.35	1.28	1.20
INMCM4	123.83	1.34	1.20
MRI_CGCM3	126.15	1.48	1.12

Table 11. Annual bias in the number of heavy precipitation days (precipitation > 10mm day^{-1}) and hot days (Tmax > 32^{o}C (90^{o}F) for the southern US regions (defined in Figure 13). Observed actual values from the GHCN and CPC datasets are shown in parentheses.

	Number of heavy precipitation days			Number of hot days		
	SW	SC	SE	SW	SC	SE
Obs	(8.5)	(23.4)	(37.5)	(55.8)	(59.5)	(40.1)
CanESM2	-6.8	-16.9	-24.3	13.8	8.0	34.7
CCSM4	2.4	-9.6	-11.1	-8.3	-5.6	-20.6
GFDL-CM3	0.1	-16.1	-23.1	-39.9	-49.4	-37.4
GFDL-ESM2G	7.8	3.0	1.3	-35.4	-31.0	-23.6
GFDL-ESM2M	8.2	3.1	2.4	-33.1	-26.4	-19.4
GISS-E2-R	18.9	7.3	11.4	-34.9	-41.3	-35.5
HadCM3	-6.3	-20.9	-29.2	5.0	8.3	-15.0
HadGEM2-CC	2.0	-0.5	1.3	7.8	-5.5	-18.2
IHadGEM2-ES	-3.5	-5.3	-1.0	9.6	-0.3	-15.8
IPSL-CM5A-LR	-3.9	-20.5	-29.0	-49.0	-35.5	-38.3
MIROC4h	1.3	-5.0	-2.5	-17.2	14.3	10.7
MIROC5	-3.2	-13.1	-13.5	-14.1	13.8	1.6
MPI-ESM-LR	2.6	-8.9	-4.3	-26.9	-19.0	-27.6
MRI-CGCM3	9.6	-4.9	-1.3	-39.6	-46.1	-38.7
MME mean	2.1	-7.7	-8.8	-18.7	-12.6	-17.4

Table 12. Annual mean RMSE for precipitation (mm day^{-1}) for each of the 17 core CMIP5 models compared with CMAP observed estimates for the North American Monsoon region 20-35^{o}N, 102.5-115^{o}W.

Model	RMSE (mm day^{-1})
BCC-CSM1-1	1.92
CCSM4	1.53
CNRM-CM5	1.29
CSIRO-Mk3	1.09
CanESM2	0.44
GFDL-CM3	1.54
GFDL-ESM2M	1.72
GISS-E2-R	1.46
HadCM3	0.63
HadGEM2-ES	0.75
INMCM4	1.11
IPSL-CM5A-LR	0.99
MIROC-ESM	1.32
MIROC5	1.58
MPI-ESM-LR	1.09
MRI-CGCM3	2.08
NorESM1-M	1.96

Table 13. CMIP5 model error statistics for the simulation of the NAM in the core region, calculated with respect to the P-NOAA observational dataset.

Model	RMSE (mm day^{-1})	Bias (%)	Lag (months)
BCC-CSM1-1	1.96	83.6	0
CanESM2	1.11	-41.6	0
CCSM4	1.24	70.7	0
CNRM-CM5	0.75	40.4	0
CSIRO	0.77	24.6	0
GFDL-CM3	1.57	74.6	1
GFDL-ESM2G	2.74	137.7	0
GFDL-ESM2M	2.48	117.8	1
GISS-E2-R	1.90	23.5	4
HadCM3	0.83	0.2	0
HadGEM2-CC	0.88	44.8	0
HadGEM2-ES	0.85	37.9	0
INMCM4	1.67	-9.7	4
IPSL-CM5A-LR	1.26	29.6	1
IPSL-CM5A-MR	0.92	1.4	1
MIROC-ESM	1.64	40.0	2
MIROC4h	1.32	65.4	0
MIROC5	1.71	91.4	0
MPI-ESM-LR	1.36	72.8	0
MRI-CGCM3	1.40	79.4	0
NorESM1-M	2.33	110.4	1
MME mean	1.46	52.10	0.71

Table 14. Error statistics for the simulation of the GPLLJ. The statistics are calculated over the regions shown in Fig. 17 and are the RMSE and the Index of Agreement (Legates and McCabe 1999; McCabe et al. 2002).

Model	RMSE				Index of Agreement			
	Intensity (vertical)	Seasonal cycle	Spatial extent	Average	Vertical structure	Seasonal cycle	Spatial extent	Average
CanESM2	0.87	0.85	0.87	0.87	0.95	0.96	0.94	0.95
CCSM4	0.91	0.88	0.86	0.88	0.95	0.96	0.94	0.95
CNRM-CM5	0.66	0.74	0.85	0.75	0.97	0.97	0.93	0.96
GFDL-ESM2M	0.90	1.01	0.85	0.92	0.93	0.93	0.92	0.93
HadGEM2-ES	1.06	0.98	0.84	0.96	0.92	0.95	0.95	0.94
MIROC5	1.12	0.65	0.88	0.88	0.91	0.98	0.93	0.94
MPI-ESM-LR	0.76	0.85	0.77	0.79	0.95	0.96	0.94	0.95
MRI-CGCM3	1.04	1.28	0.92	1.08	0.90	0.89	0.90	0.90

Table 15. Biases in CMIP5 model Arctic sea ice extent and thickness**.** Biases are based on the ensemble mean for each model that has more than one ensemble member and computed relative to the observed value. September extent bias in 10^6 km^2.. March ice thickness bias in meters.

Model	September extent bias 10^6 km^2	March thickness bias m
BCC-CSM1-1	-0.439	-0.05
CanCM4	-1.881	-0.76
CanESM2	-2.221	-0.44
CCSM4	0.537	0.08
CESM1-CAM5	0.698	0.11
CNRM-CM5	-0.638	-0.27
CSIRO-MK6	3.952	0.31
FGOALS-s2	2.309	-0.06
GFDL-CM3	0.231	-0.37
GISS-E2-H	-2.979	-
GISS-E2-R	-2.653	-0.31
HadCM3	-0.772	-0.20
HadGEM2	-1.845	-1.06
HadGEM2-CC	-0.035	-0.52
HadGEM2-ES	-1.318	-0.90
INMCM4	-1.468	-0.04
IPSL-CM5A-LR	0.708	0.02
IPSL-CM5A-MR	-0.718	-0.32
MIROC-ESM-CHEM	-0.465	-0.81
MIROC-ESM	-0.783	-0.85
MIROC4h	-1.673	-0.71
MIROC5	-0.918	-0.04
MPI-ESM-LR	0.070	-0.57
MRI-CGCM3	-0.931	-0.16
NorESM1-M	1.205	-0.30

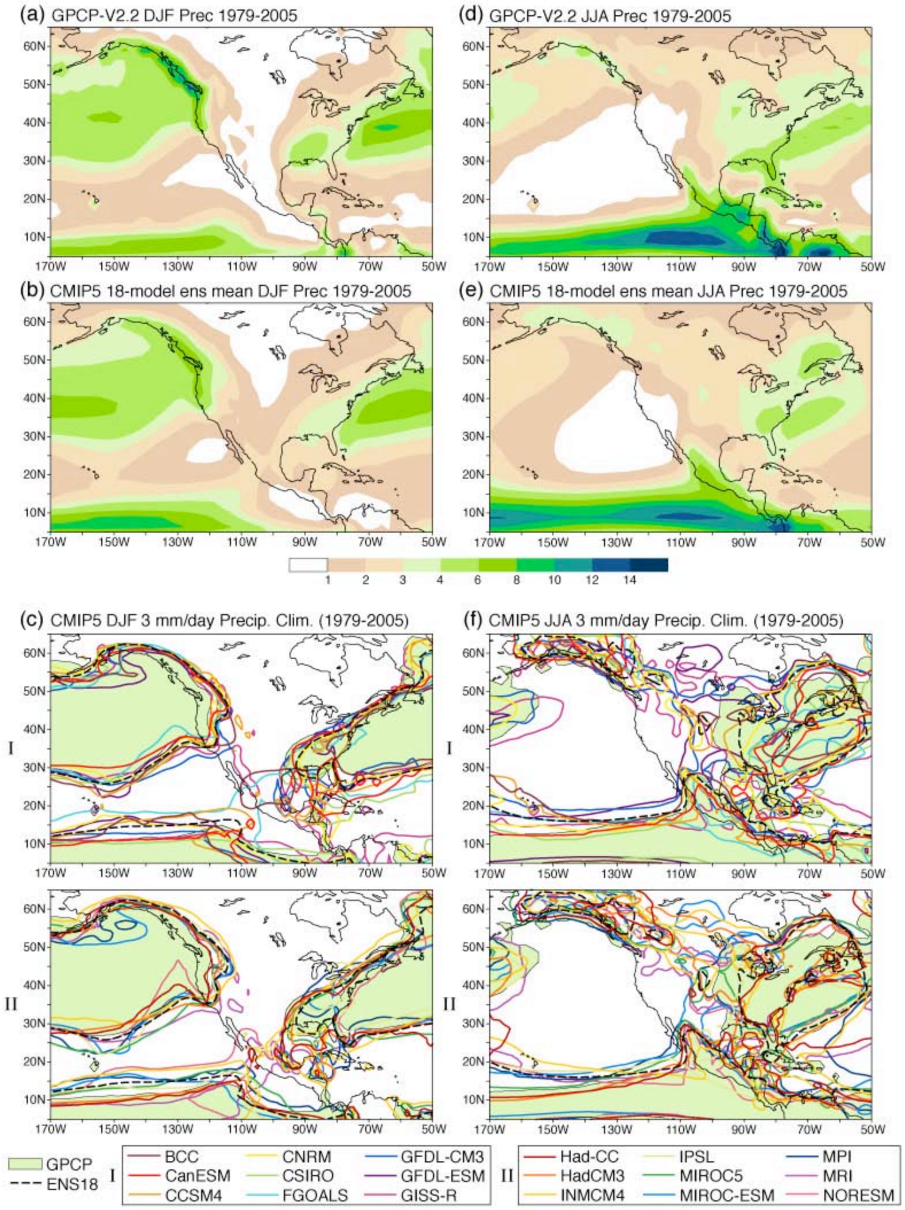

Figure 1. Precipitation climatology for (left) December-February and (right) June-August (1979-2005). a) GPCP estimate of observed precipitation for DJF. b) MME mean over the 18 models for DJF; for models with multiple runs, all runs are averaged before inclusion in the multi-model ensemble. c) Comparison of individual models to observations using the 3 mm day^{-1} contour as an index of the major precipitation features: half the models are shown in each of sub-panel I and II with the legend giving the color-coding for the models in each. Shading shows the regions where GPCP exceeds 3 mm day^{-1}; a model with no error would have its contour fall exactly along the edge of the shaded region. d)-f) As in a)-c), respectively, except for JJA.

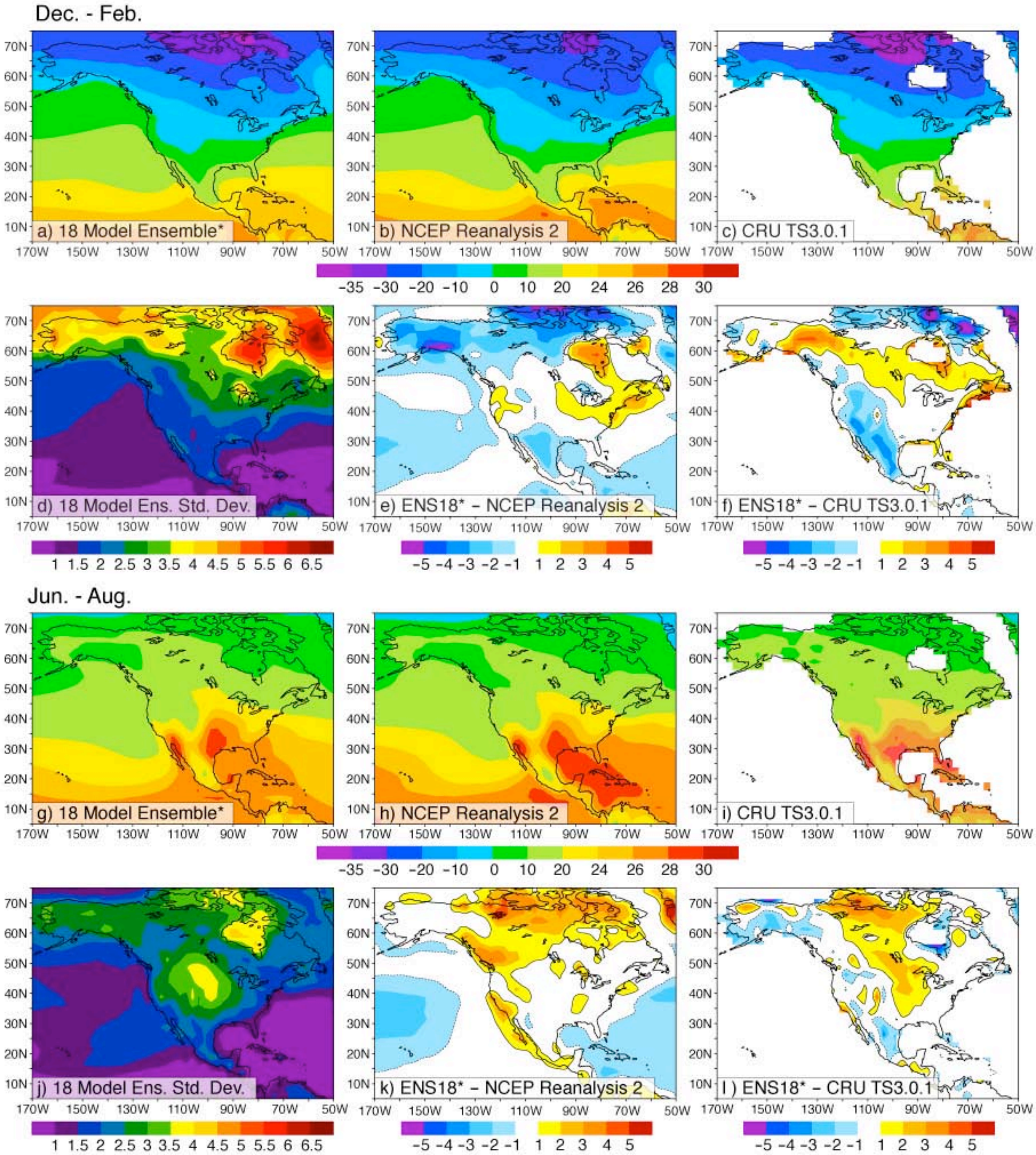

Figure 2. Surface air temperature climatology for (top) December-February and (bottom) June-August (1979-2005). a) MME mean (over the 17 core models plus FGOALS-s2) for DJF. b) NCEP-DOE Reanalysis 2 estimate of observed surface air temperature climatology for DJF. c) As in b) but for CRU. d) Standard deviation of surface air temperature among the 18 model DJF climatological values at each point. e) Difference between the MME mean climatology in a) and the NCEP-DOE Reanalysis 2. f) As in e) but for CRU. g)-l) As in a)-f) but for JJA.

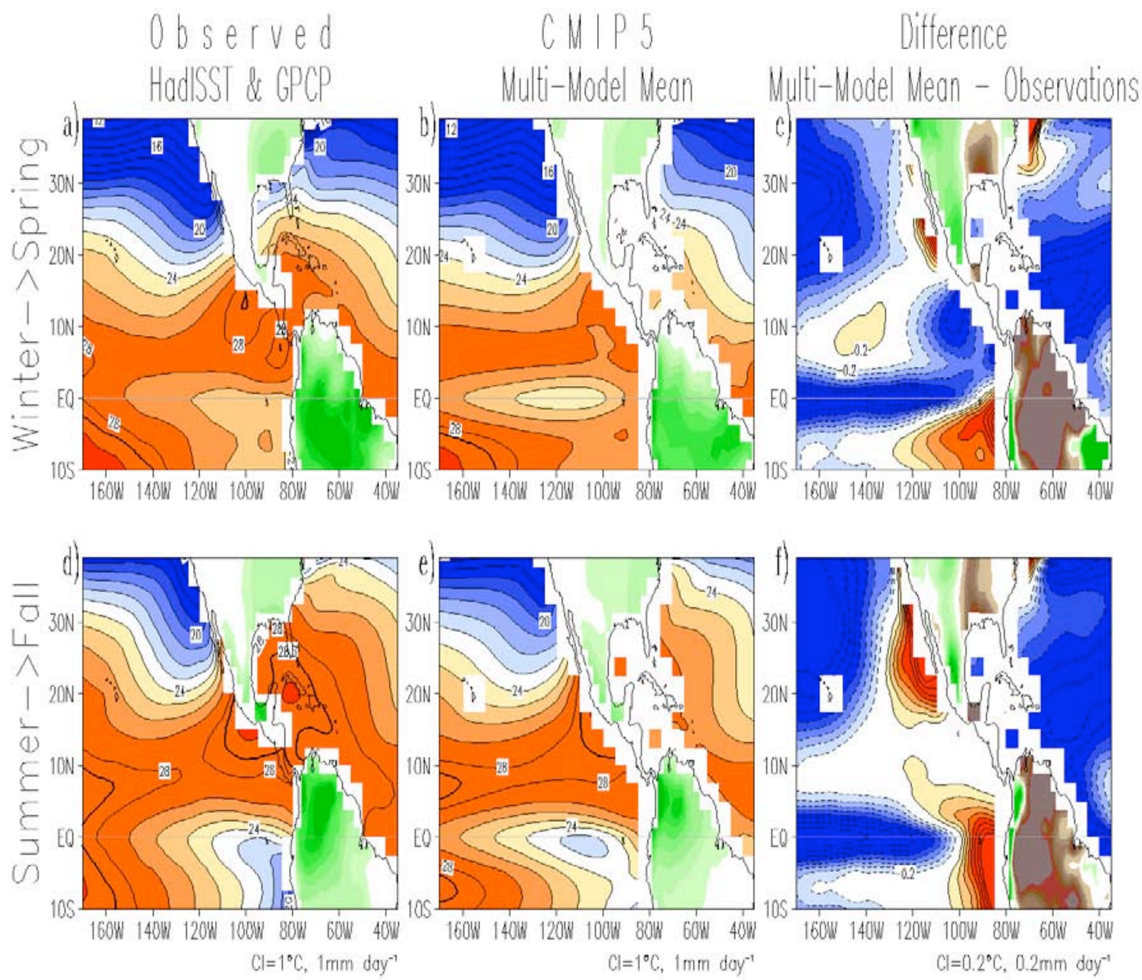

Figure 3. Climatological sea surface temperature and precipitation in observations from HadISSTv1.1 and GPCPv2.2 data sets, and historical simulations from 17 CMIP5 models for 1979-2004. (a) Observations, (b) MME mean, and (c) MME mean minus observations, for winter-to-spring (December to May). (d-f) as in (a-c) but for summer-to-fall (April to November). Temperatures are shaded blue/red for values equal or lower/larger than 23/24°C; the thick black line highlights the 28.5°C isotherm as indicator of the Western Hemisphere Warm Pool. Precipitation is shaded green for values equal or larger than 2 mm day^{-1}. Contour intervals are 1°C and 1 mm day^{-1} for the mean values and 0.2^{o}C and 2 mm day^{-1} for the differences. SST/precipitation fields have been regridded to a common 5°×2.5°/2.5°×2.5° grid.

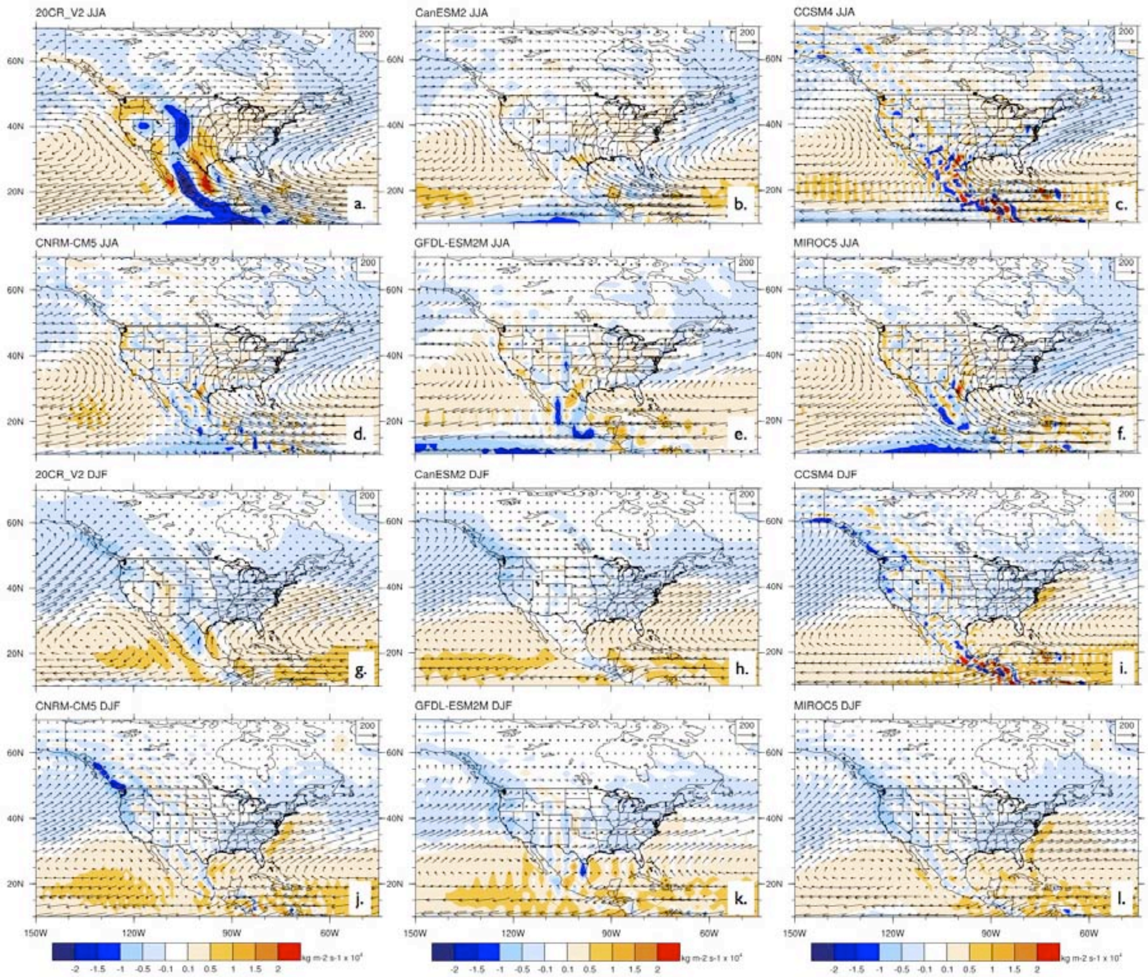

Figure 4. Vertically integrated moisture transport (vectors) and its divergence (contours) for the 20CR reanalysis (a,g) and five CMIP5 models for mean JJA (b-f) and DJF (h-l) for 1981-2000. Vertically integrated moisture transport is computed to 500 hPa using 6-hourly data from the 20CR and one realization each from the historical experiments for CanESM2, CCSM4, CNRM-CM5, GFDL-ESM2M, and MIROC5 models.

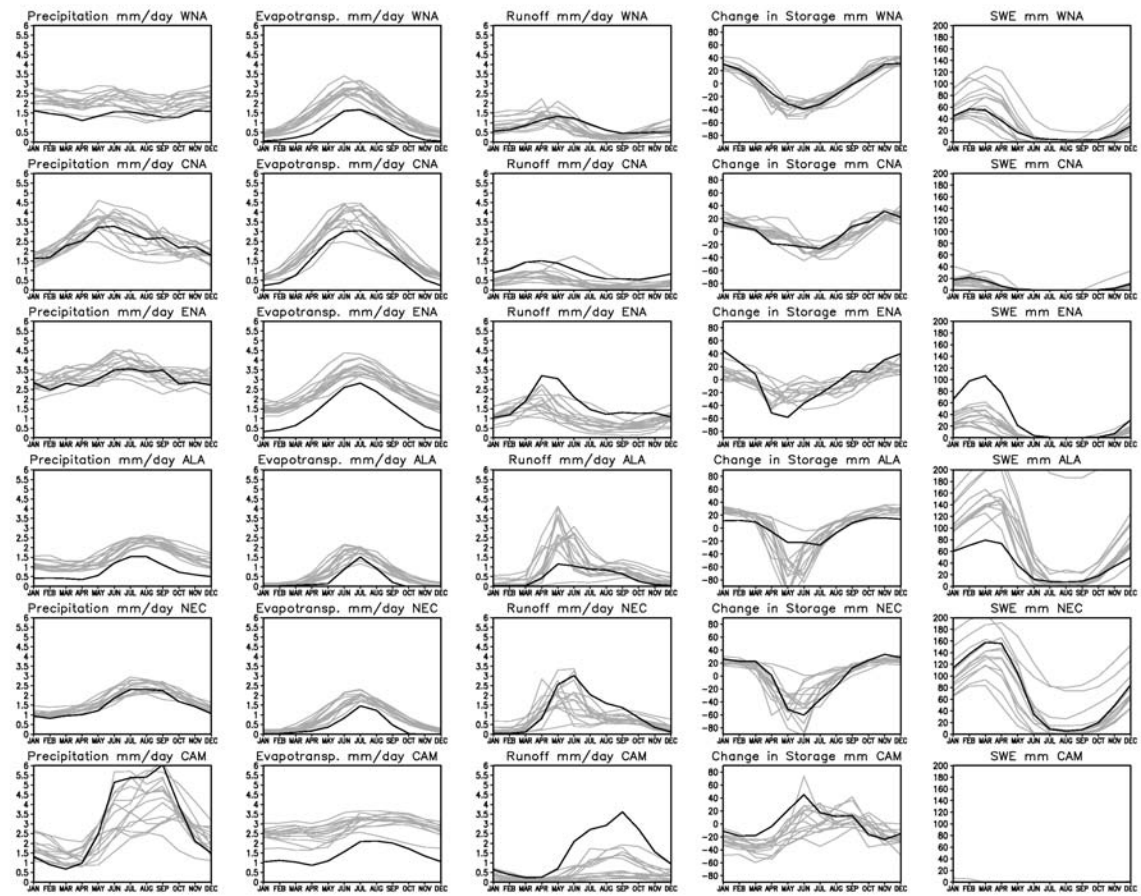

Figure 5. Mean seasonal cycle (1979-2005) of North American regional land water budget components for 12 CMIP5 models (CanESM2, CSIRO-Mk3-6-0, GFDL-ESM2G, GISS-E2-H, GISS_E2-R, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC-ESM, MIROC-ESM-CHEM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M) compared to the average of the two off-line LSM simulations (VIC and GLDAS2 Noah). Regions are Western North America (WNA), Central North America (CNA), Eastern North America (ENA), Alaska and Western Canada (ALA), Northeast Canada (NEC), and Central America (CAM) as modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3.

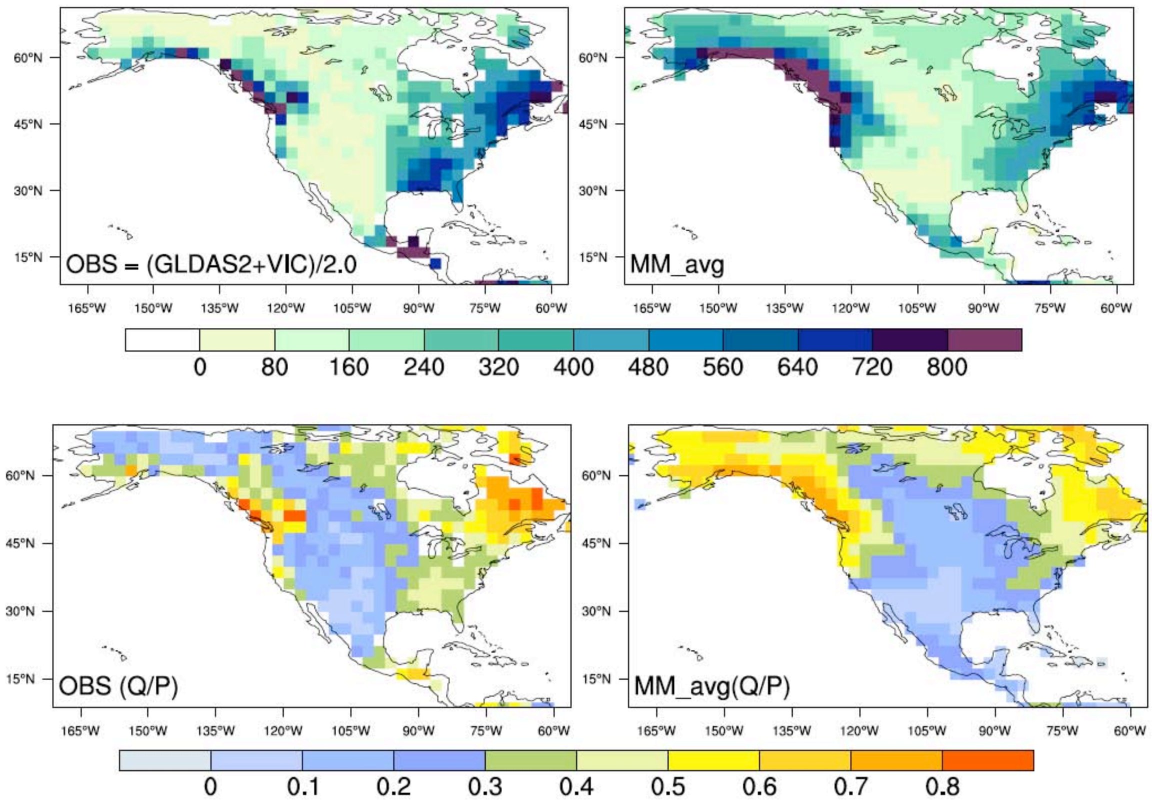

Figure 6. Mean annual runoff (mm/year) (top) and runoff (Q/P) ratio (bottom) for 1979-2004 from observations (Q from VIC and GLDAS2 Noah, and P from GPCP) and the multi-model average from 15 CMIP5 climate models (BCC-CSM1-1, CanESM2, CCSM4, CNRM-CM5, CSIRO-Mk3-6-0, GFDL-CM3, GFDL-ESM2M, GISS-E2-R, HadCM3, INMCM4, MIROC5, MIROC-ESM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M).

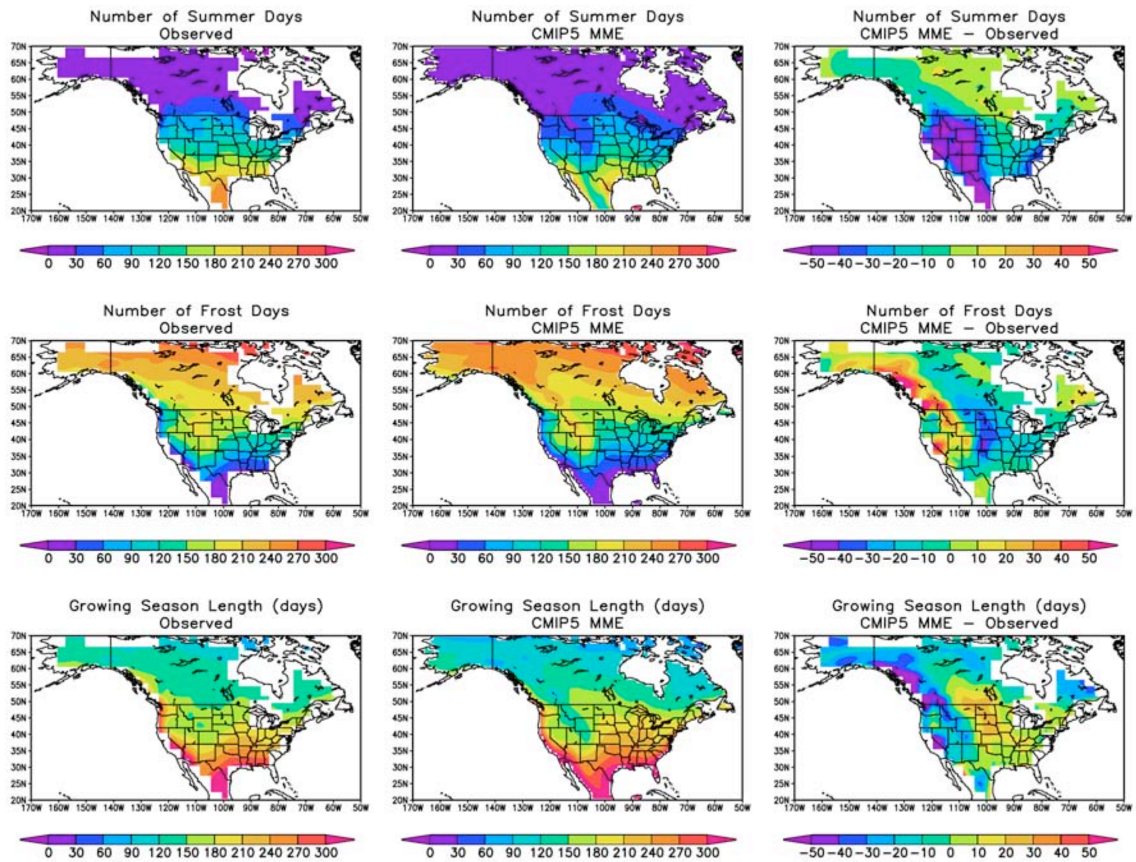

Figure 7. Comparison of biophysical indicators between observations and the CMIP5 ensemble. Biophysical indicators are (top row) number of summer days, (middle row) number of frost days, and (bottom row) growing season length averaged over 1979-2005. Left column shows the observations (left column) from the HadGHCND dataset; middle column is the multi-model ensemble mean of the 17 core models, and the right column is their difference (MME – obs.). The frequencies are calculated on the model grid and then interpolated to 2.0 degree resolution for comparison with the observational estimates.

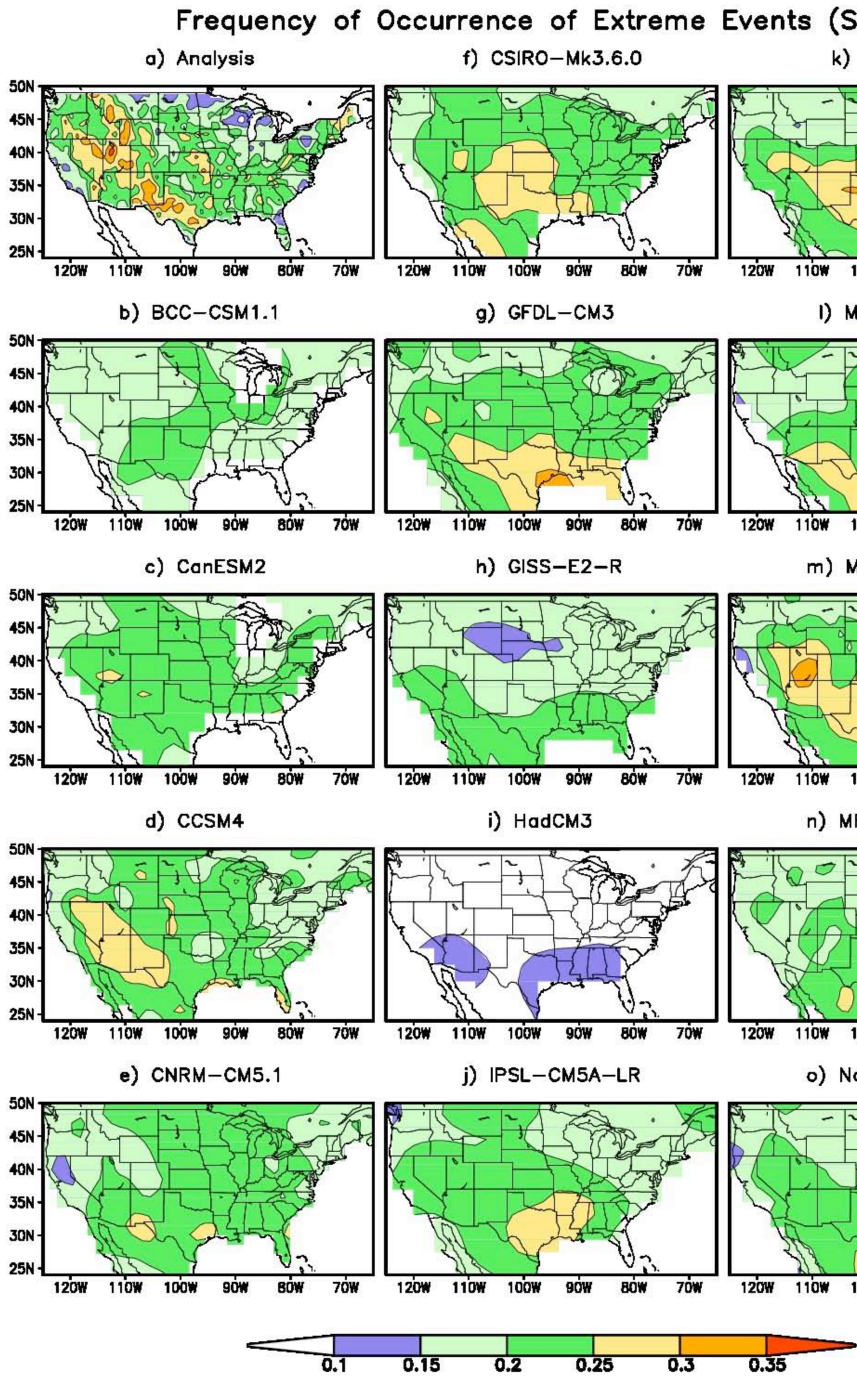

Figure 8. The frequency of occurrence of persistent extreme precipitation events defined by SPI6 averaged over positive and negative events for (a) observed precipitation based on the CPC and UW datasets, (b) BCC-CSM1-1, (c) CanESM2, (d) CCSM4, (e) CNRM-CM5.1, (f) CSIRO-Mk3.6.0, (g) GFDL-CM3, (h) GISS-E2-R, (i) HadCM3, (j) IPSL-CM5A-LR, (k) MIROC5, (l) MIROC-ESM, (m) MPI-ESM-LR, (n) MRI-CGCM3 and (o) NorESM1-M. The HadCM3 and HadGEM2-ES results are similarly weak and so the former are shown only. Each data set is treated as one member of the ensemble.

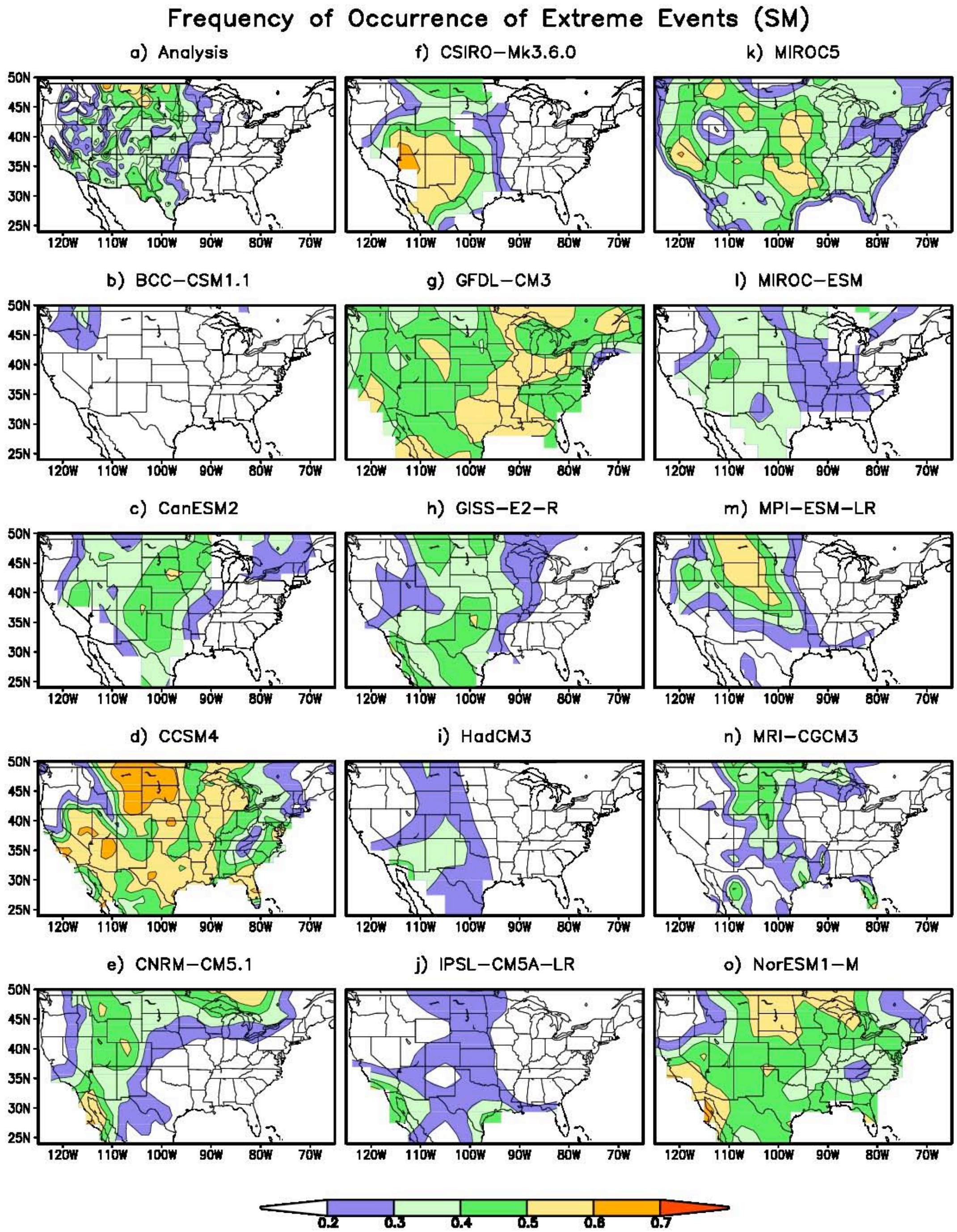

Figure 9. Same as Figure 8 but for persistent soil moisture events. Estimates of observed soil moisture are taken from the multi-model NLDAS-UW dataset.

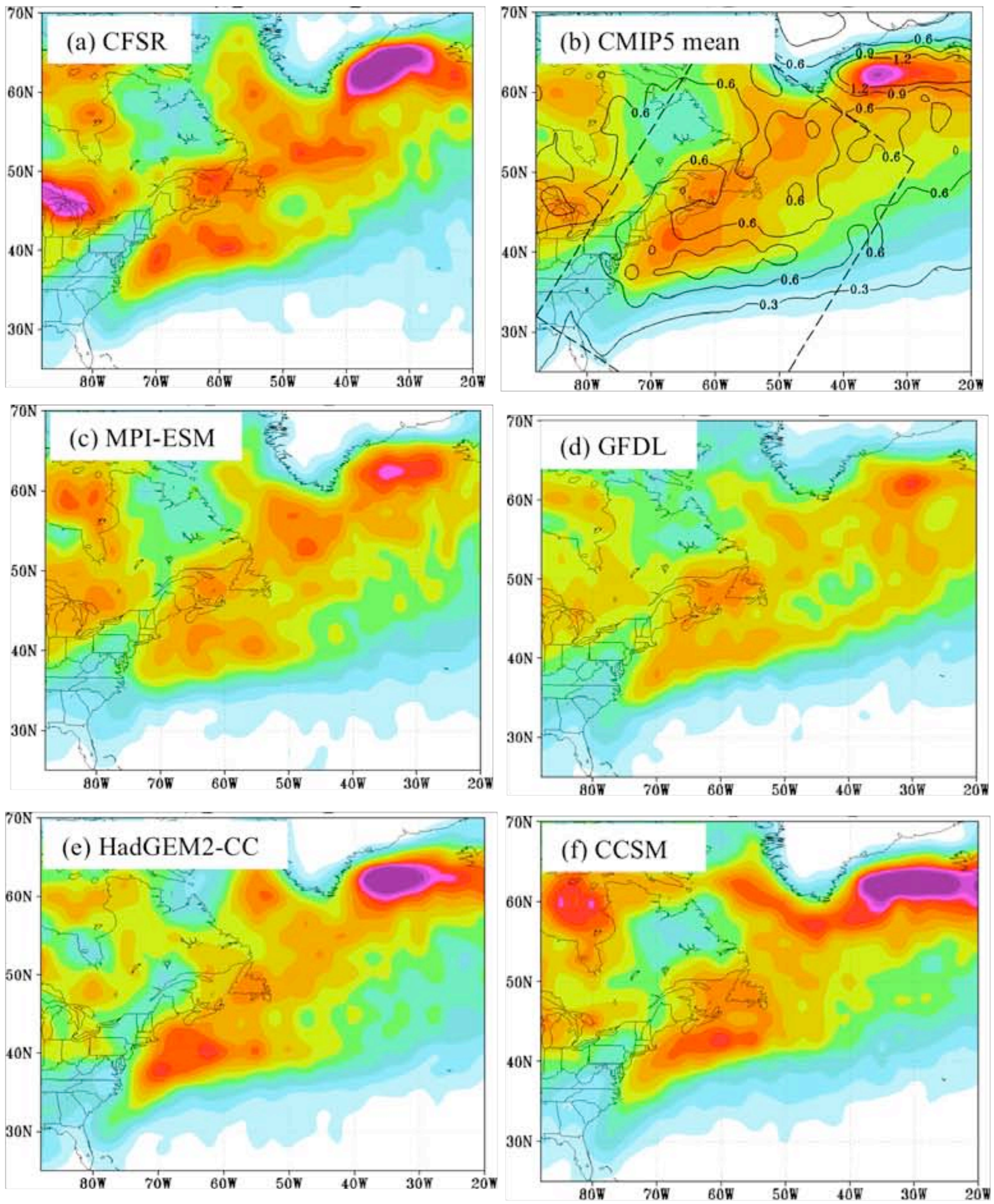

Figure 10. (a) Cyclone density for the CFSR analysis showing the number of cyclones per cool season (November to March) per 50,000 km^2 for 1979-2004. (b) Same as (a) except for the mean (shaded) and spread (contoured every 0.3) of 15 CMIP5 models ordered from higher to lower spatial resolution: CanESM2, EC-EARTH, MRI-CGCM3, CNRM-CM5, MIRCO5, HadGEM2-ES, HadGEM2-CC, INMCM4, IPSL-CM5A-MR, MPI-ESM-LR, NorESM1-M, GFDL-ESM2M, IPSL-CM5A-LR, BCC-CSM1, MIROC-ESM-CHEM. Same as (a) except for the (c) MPI-ESM-LR, (d) GFDL-ESM2M, (e) HadGEM2-CC, and (f) CCSM4 models.

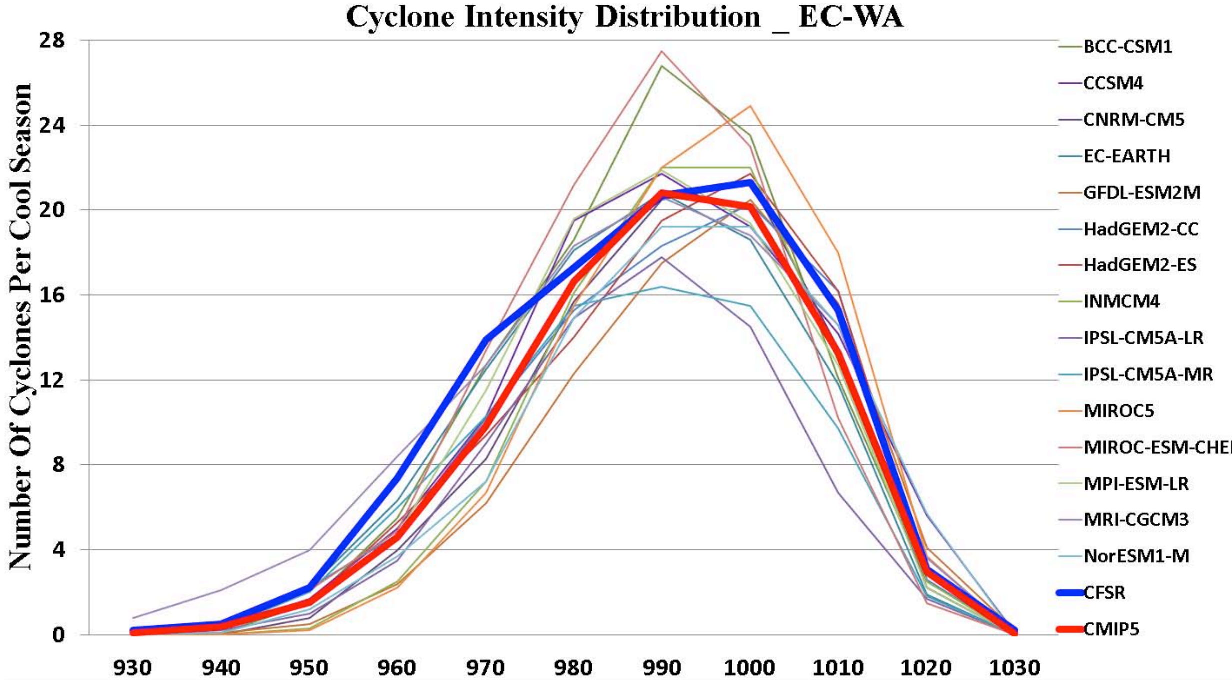

Figure 11. Number of cyclone central pressures at their maximum intensity (minimum pressure) for the 1979-2004 cool seasons within the dashed box region in Fig. 10 for a 10 hPa range centered every 10 hPa showing the CFSR (bold blue), (b) CMIP5 MME mean (bold red), and individual CMIP5 models.

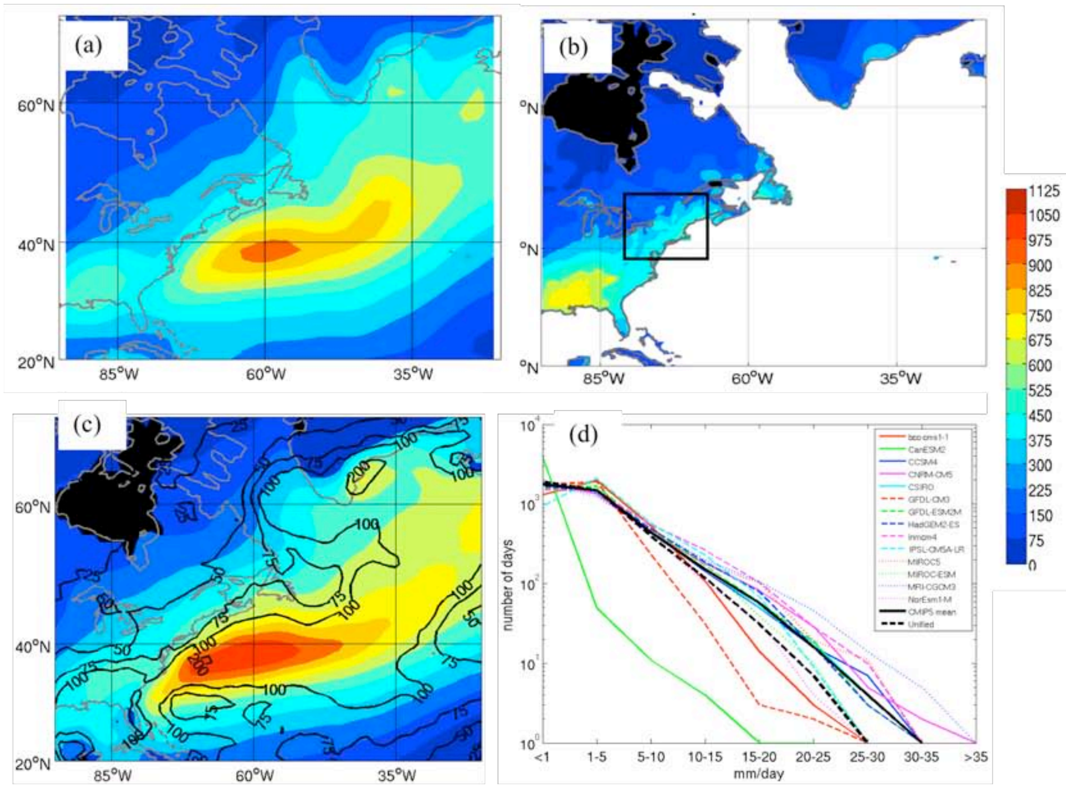

Figure 12. (a) CPC merged precipitation analysis at 2.5 deg resolution showing cool seasonal average precipitation (shaded every 75 mm) for the 1979-2004 cool seasons (November – March). (b) Same as (a) except for the CPC Unified precipitation at 0.5 deg resolution. (c) Same as (a) except for the mean of 14 of the 17 CMIP5 members listed in (d) and spread (in mm). (d) Number of days that the daily average precipitation (in mm/day) for the land areas in the black box in (b) occurred within each amount bin for select CMIP5 members, CMIP5 mean, and the CPC Unified.

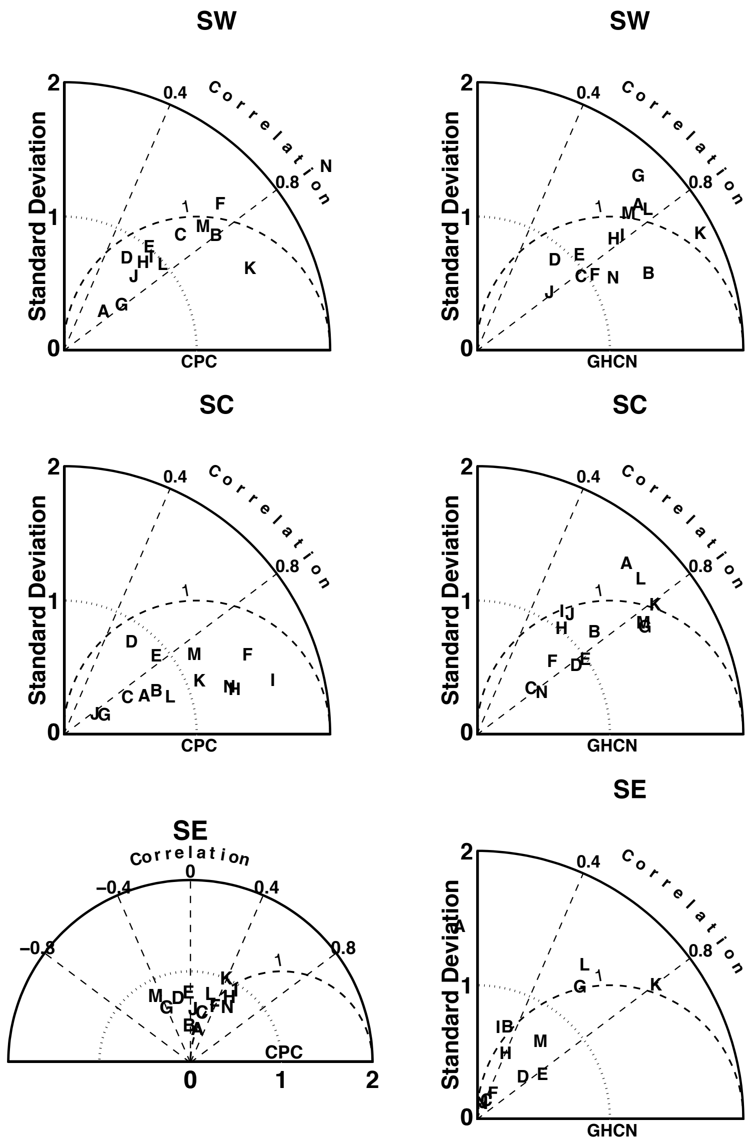

Figure 13. Comparison of precipitation and temperature extremes for southern US regions between the CMIP5 models and CPC and GHCN observations, respectively. (left column) Taylor diagram of the spatial pattern of annual number of days when precipitation > 10mm day^{-1} over the southwest (SW), south central (SC) and southeastern (SE) US. (right column) Taylor diagram of the spatial pattern of annual number of days when Tmax > 32^{o}C (90F) for the three regions. The standard deviations have been normalized relative to the observed values. (A: CanESM2, B: CCSM4, C: GFDL-CM3, D: GFDL-ESM2G, E: GFDL-ESM2M, F: GISS-E2-R, G: HadCM3, H: HadGEM2-CC, I: HadGEM2-ES, J: IPSL-CM5A-LR, K: MIROC4h, L: MIROC5, M: MPI-ESM-LR, N: MRI-CGCM3). Observations are from the CPC dataset. SW is defined as the contiguous US south of 40^{o}N between 125^{o}W and 110^{o}W; SC is the contiguous US south of 40^{o}N between 110^{o}W and 90^{o}W; SE is the contiguous US south of 40^{o}N between 90^{o}W and 70^{o}W.

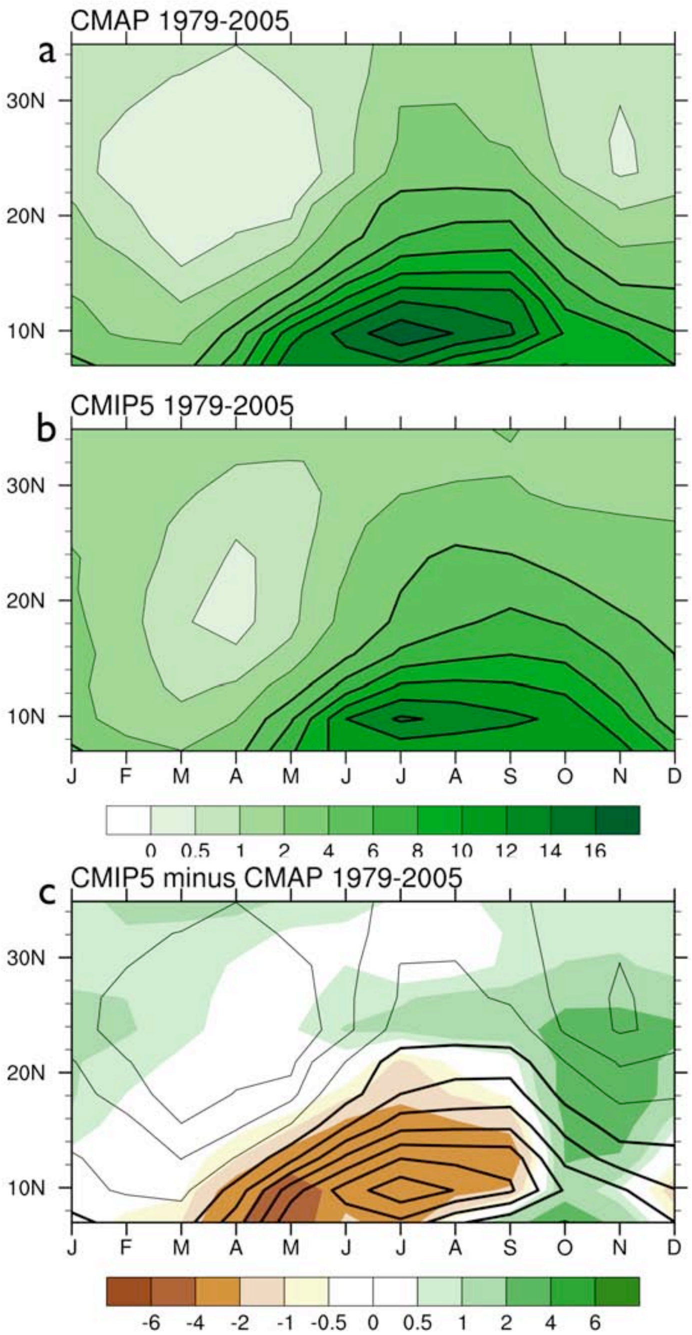

Figure 14. Average monthly precipitation for 1979-2005 shown by latitude in the North American monsoon region (longitudes 102.5 to 115W) from the CMAP observational estimate (a), the MME mean for the 17 core CMIP5 models (b) and their difference (c), all in units of mm day^{-1}.

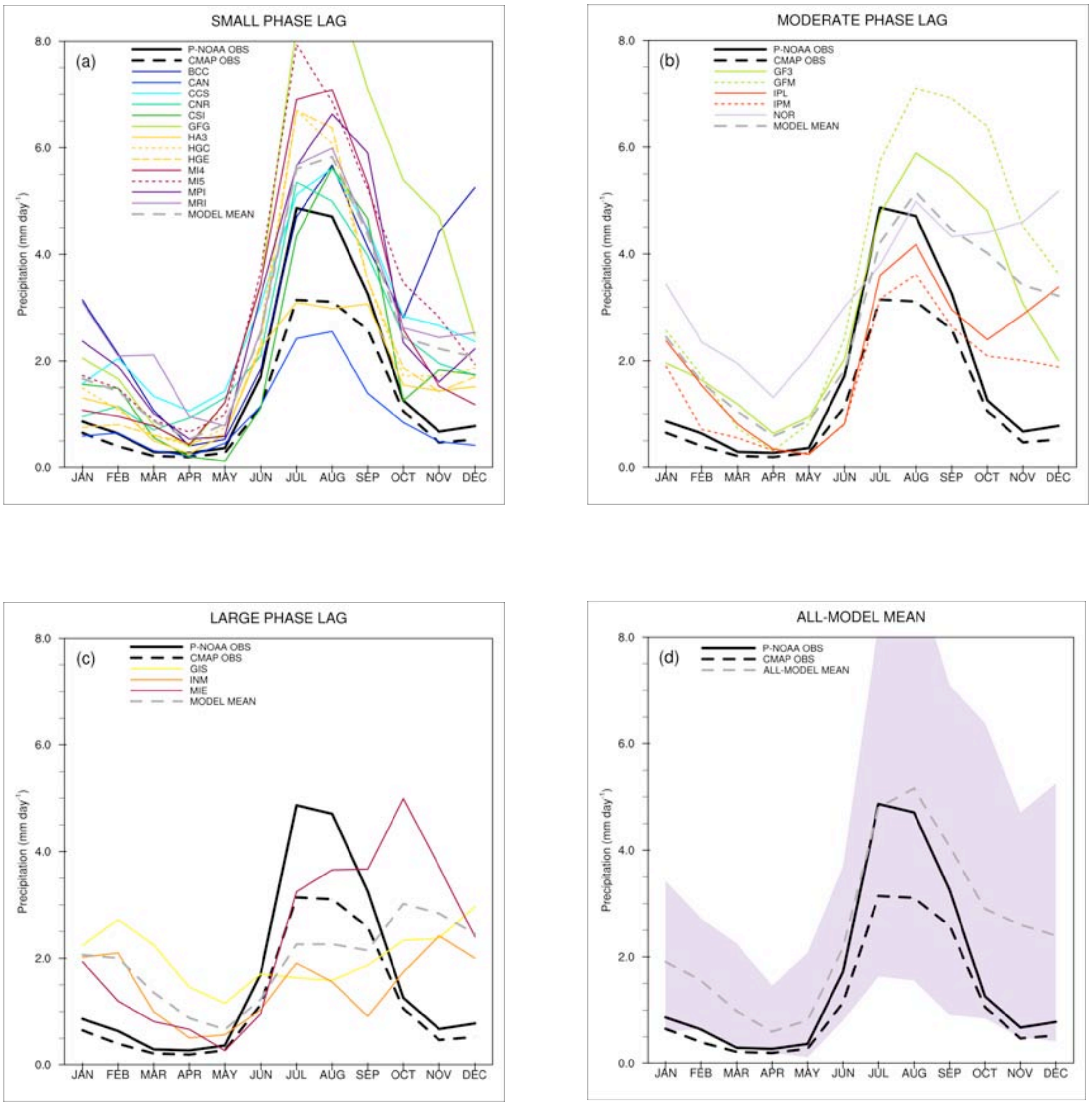

Figure 15. Annual cycle in rainfall for the NAM region for the historical (1979-2005) period of 21 CMIP5 models compared to the P-NOAA AND CMAP observational datasets for (a) small (phase error = 0), (b) moderate (phase error = 1), (c) large (phase error = 2-4) phase errors, and (d) all models.

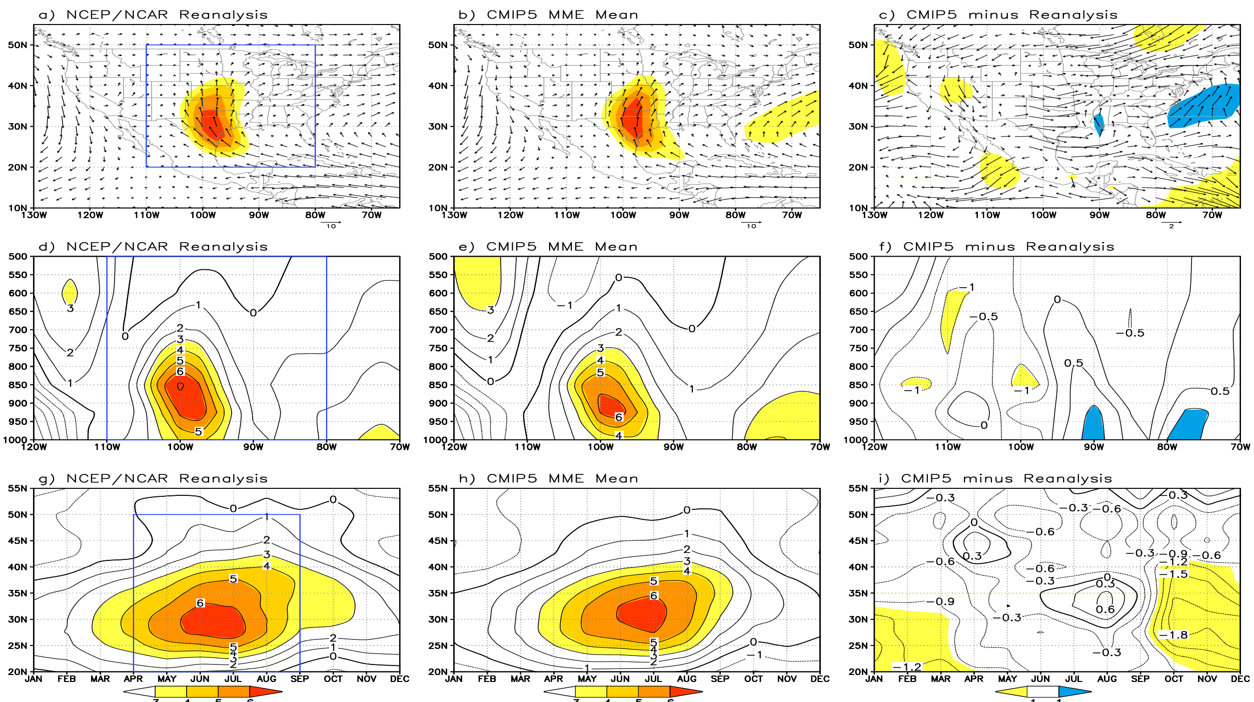

Figure 16. (a)-(c) Averaged summer 925hPa wind during 1971-2000 for NCEP-NCAR reanalysis, eight-model CMIP5 ensemble mean for the same period, and the reanalysis minus MME mean, respectively. (d)-(f) Lower troposphere mean vertical profile of meridional wind averaged over 95˚-100˚W for the reanalysis, MME mean, and the reanalysis minus MME mean, respectively. (g)-(i) Seasonal cycle of the 925hPa meridional wind averaged over 27.5º-32.5ºN for the reanalysis, MME mean, and the reanalysis minus MME mean. All units are m s^{-1}. Shading indicates wind speeds greater than 3.0 m s^{-1} in the figures of the first and second columns and wind speeds greater than 1.0m s^{-1} in the figures of the third column.

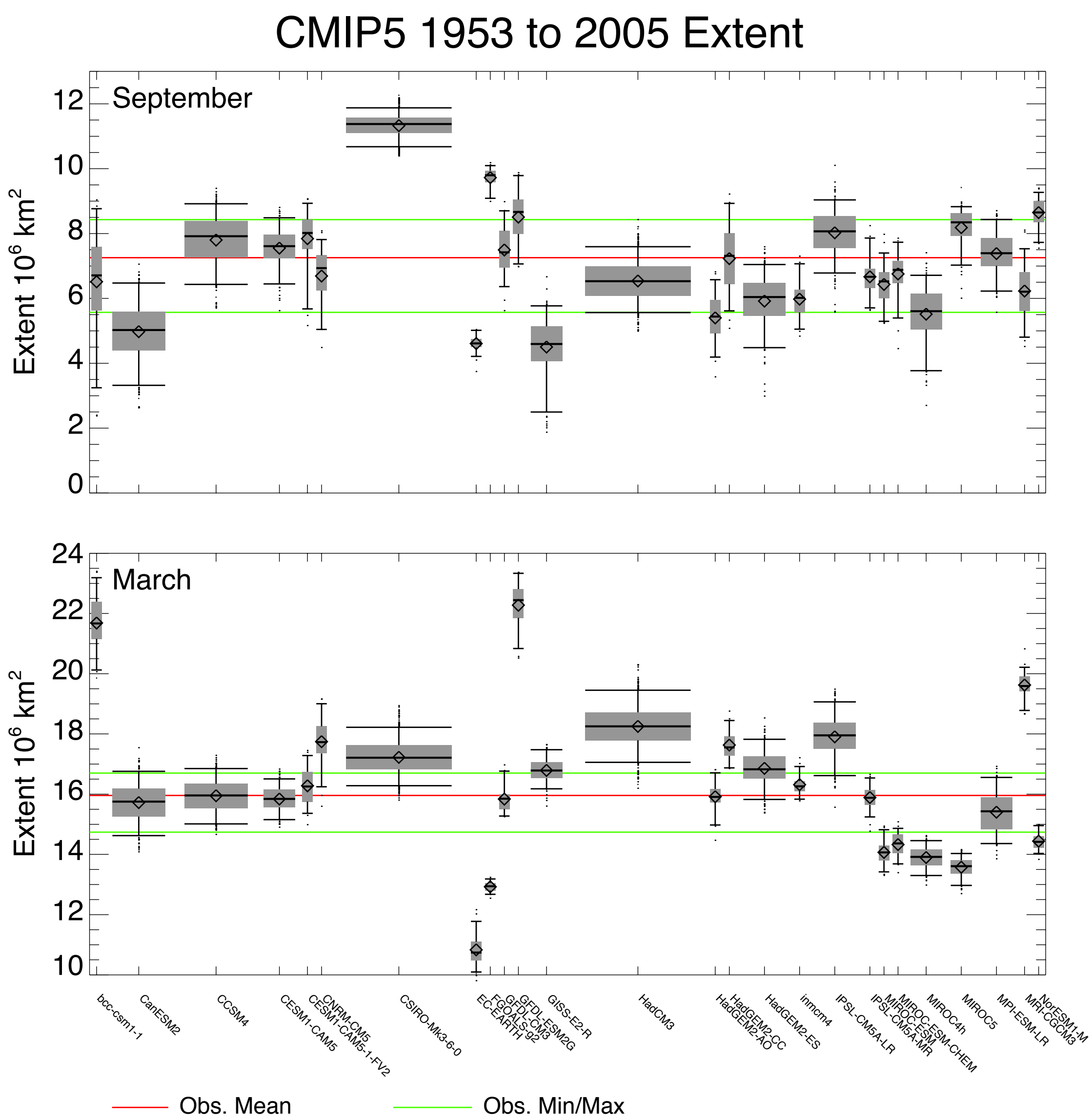

Figure 17. September and March sea ice extent from 26 CMIP5 models compared to observations from the NSIDC from 1953 to 2005. For each model, the boxes represent inter-quartile ranges (25th to 75th percentiles). Median (50th percentile) extents are shown by the thick horizontal bar in each box. The width of each box corresponds to the number of ensemble members for that model. Whiskers (vertical lines and thin horizontal bars) represent the 10th and 90th percentiles. Mean monthly extents are shown as diamonds. Corresponding mean, minimum and maximum observed extends are shown as red and green lines, respectively.

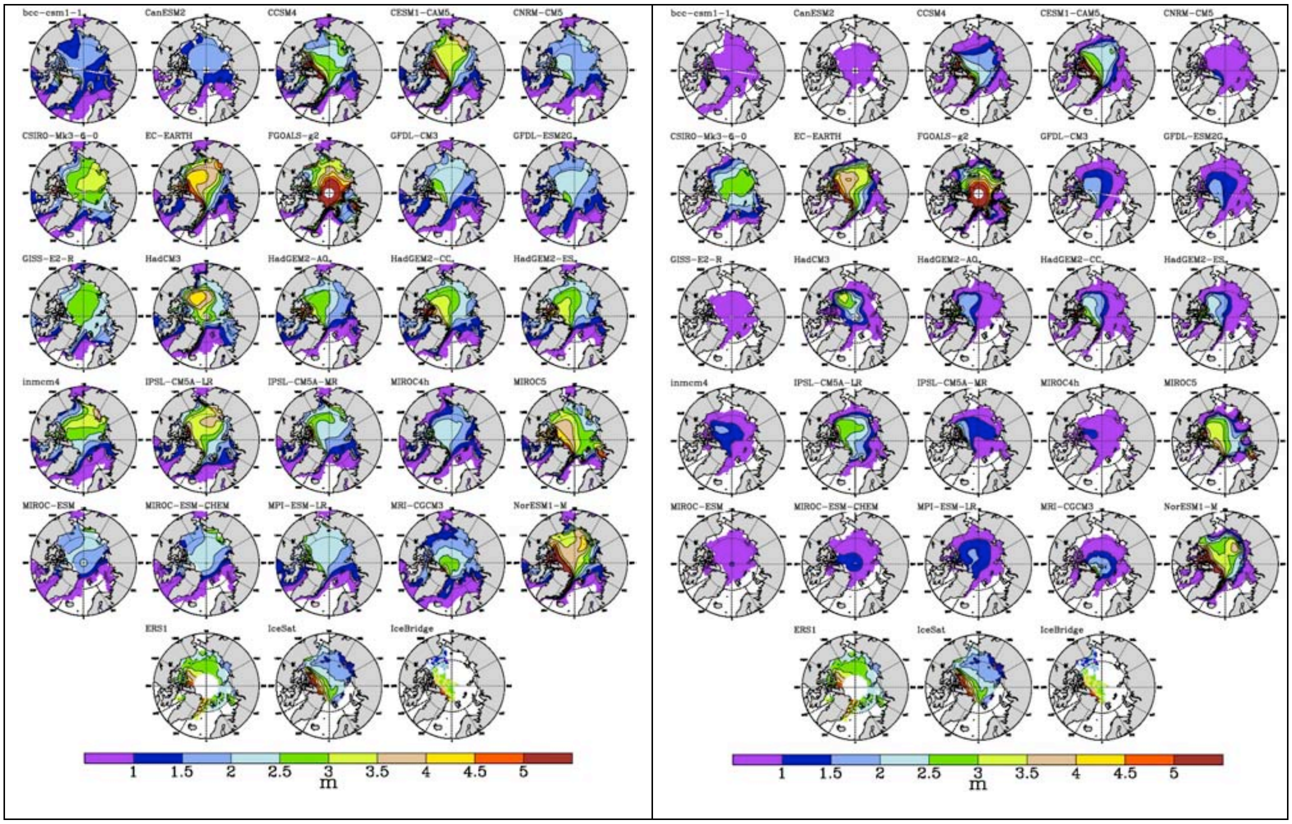

Figure 18. March (left) and September (right) ice thickness (m) for 26 CMIP5 models averaged over 1993-2005 versus satellite and airborne observations for ERS1/2 (1993-2001), ICESat (2003-2009) and IceBridge (2009-2012).

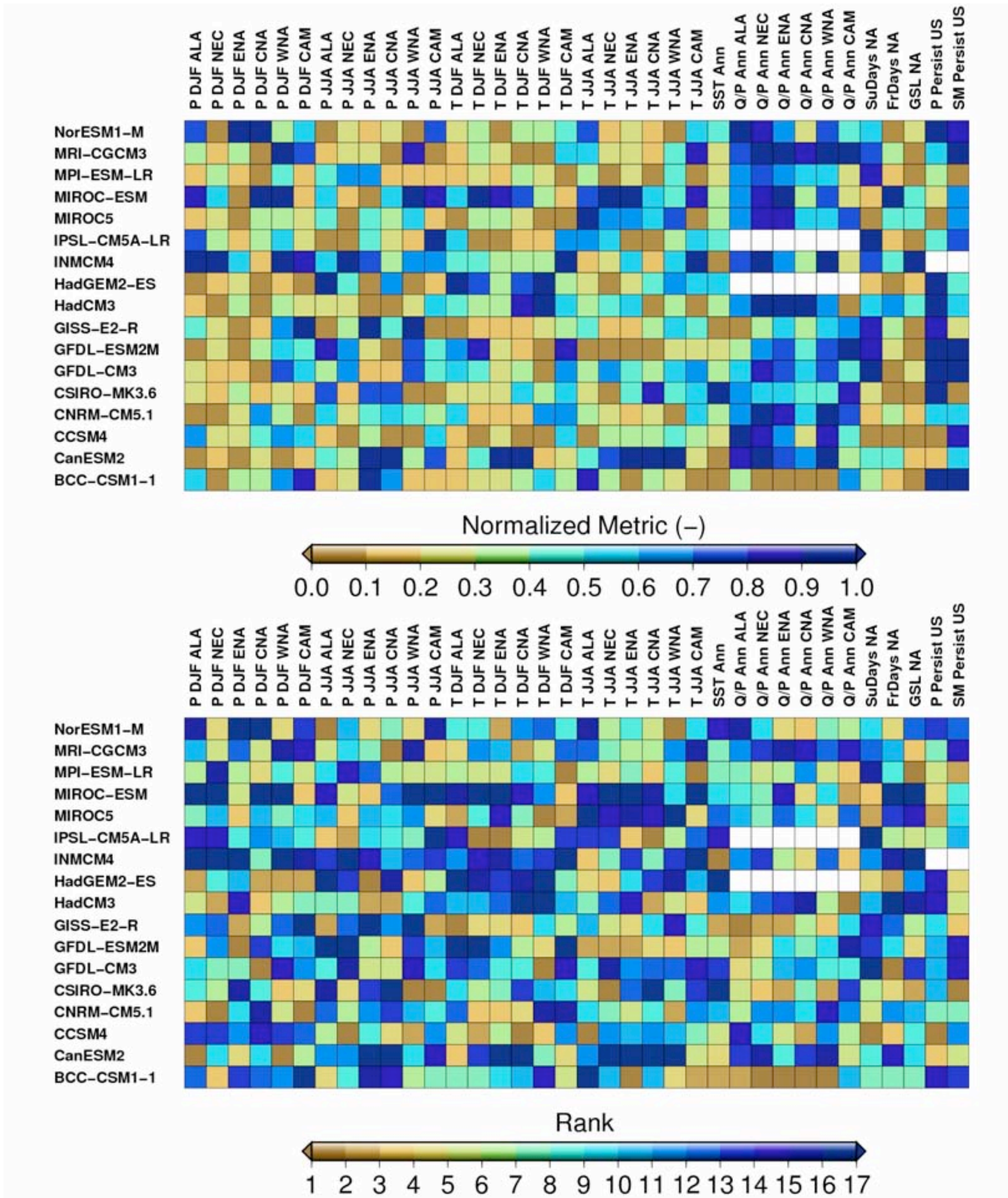

Figure 19. Comparison of CMIP5 models across a set of continental performance metrics based on bias values given in Tables 3-8. (top) Biases normalized relative to the range of bias values across models, with lower values indicating lower bias. (bottom) Models ranked according to bias values, with 1 indicating the model with the lowest bias and 17 the model with the highest bias. Results for models without available data are indicated in white. The bias metrics shown (in order from left to right) are for regional precipitation (P) for DJF and JJA, regional temperature (T) for DJF and JJA, annual SSTs for surrouding oceans (see Figure 3), annual runoff ratios (Q/P), the annual number of summer days (SuDays), frost days (FrDays) and growing season length (GSL), and east-west gradient in the number of persistent precipitation (P Persist) and soil moisture (SM Persist) events.

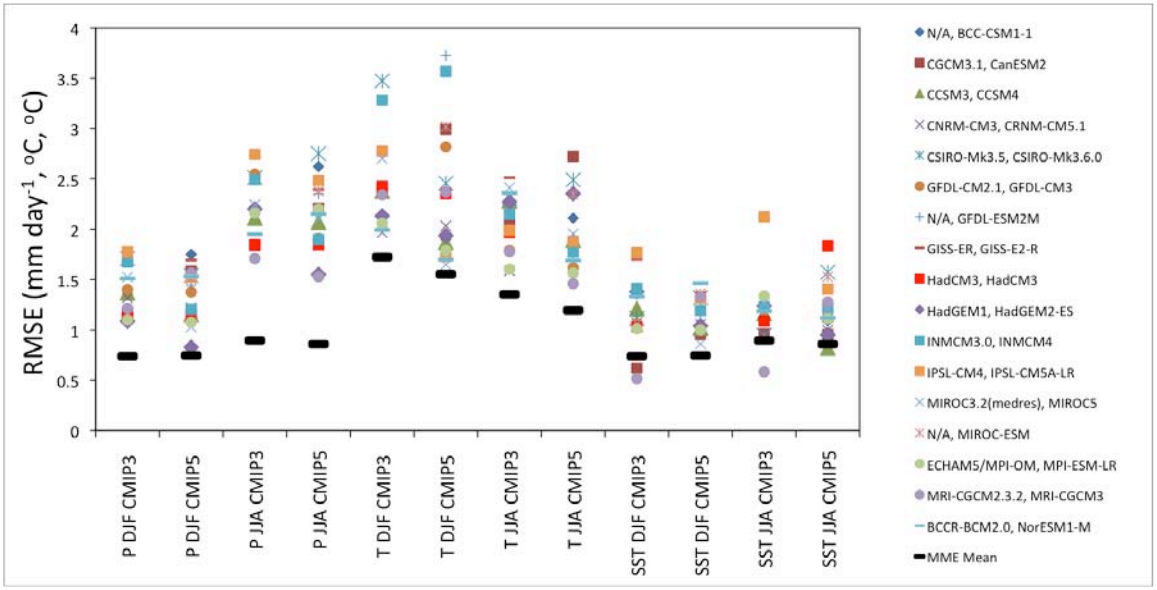

Figure 20. Comparison of CMIP5 and CMIP3 model performance for seasonal (DJF and JJA) precipitation (P), surface air temperature (T) and SST. Results are shown as RMSE values calculated for 1971-1999 relative to the GPCP, CRU and HadISST observational datasets. Precipitation and temperature RMSE values are calculated over North America (130-60W, 0-60N) and SST RMSE values are calculated over neighboring oceans (170-35W, 10S-40N). The core set of CMIP5 models and their equivalent CMIP3 models where available (otherwise indicated by N/A) are shown. The MME mean values are also shown.